Jacob Fraden

Handbook of Modern Sensors

Physics, Designs, and Applications

Fourth Edition

Springer

VISIT…

Handbook of Modern Sensors

Fourth Edition

Jacob Fraden

Handbook of Modern Sensors

Physics, Designs, and Applications

Fourth Edition

Springer

Jacob Fraden
jacob@fraden.com

ISBN 978-1-4419-6465-6 e-ISBN 978-1-4419-6466-3
DOI 10.1007/978-1-4419-6466-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2010932807

© Springer Science+Business Media, LLC 2010
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Since publication of the previous, the 3rd edition of this book, the sensor technologies have made a remarkable leap ahead. The sensitivity of the sensors became higher, the dimensions – smaller, the selectivity – better, and the prices – lower. What have not changed, are the fundamental principles of the sensor design. They still are governed by the laws of Nature. Arguably one of the greatest geniuses ever lived, Leonardo Da Vinci had his own peculiar way of praying. It went like this, "*Oh Lord, thanks for Thou don't violate Thy own laws*." It is comforting indeed that the laws of Nature do not change with time, it is just that our appreciation of them becomes refined. Thus, this new edition examines the same good old laws of Nature that form the foundation for designs of various sensors. This has not changed much since the previous editions. Yet, the sections that describe practical designs are revised substantially. Recent ideas and developments have been added, while obsolete and less important designs were dropped.

This book is about devices commonly called *sensors*. The invention of a microprocessor has brought highly sophisticated instruments into our everyday life. Numerous computerized appliances, of which microprocessors are integral parts, wash clothes and prepare coffee, play music, guard homes, and control room temperature. Sensors are essential components in any device that uses a digital signal processor. The processor is a device that manipulates binary codes generally represented by electric signals. Yet, we live in an analog world, where such devices function among objects that are mostly not digital. Moreover, this world is generally not electrical (apart from the atomic level). Digital systems, however complex and intelligent they might be, must receive information from the outside world. Sensors are the interface devices between various physical values and electronic circuits that "understand" only a language of moving electrical charges. In other words, sensors are eyes, ears, and noses of silicon chips.

In the course of my engineering work, I often felt a strong need for a book which would combine practical information on diversified subjects related to the most important physical principles, design and use of various sensors. Surely, I could find almost all I had to know by surfing Internet or browsing library bookshelves in search for texts on physics, electronics, technical magazines, manufacturer's

catalogues and websites. However, the information is scattered over many publications, and almost every question I was pondering required substantial research work. Little by little, I have been gathering practical information on everything, which in anyway was related to various sensors and their applications to scientific and engineering measurements. Soon, I realized that the information I collected might be quite useful to more than one person. This idea prompted me to write this book and this 4th edition is the proof that I was not mistaken.

In setting my criteria for selecting various sensors for the new edition, I attempted to keep the scope of this book as broad as possible, opting for many different designs described briefly (without being trivial, I hope), rather than fewer treated in greater depth. This volume attempts (immodestly perhaps) to cover a very broad range of sensors and detectors. Many of them are well known, but describing them is still useful for students and those who look for a convenient reference. It is the author's intention to present a comprehensive and up-to-date account of the theory (physical principles), design, and practical implementations of various (especially, the newest) sensors for scientific, industrial, and consumer applications. The topics included in the book reflect the author's own preferences and interpretations. Some may find a description of a particular sensor either too detailed or too broad or, on the contrary, too brief. In most cases, the author tried to strike a balance between a detailed description and simplicity of coverage.

It is clear that one book cannot embrace the whole variety of sensors and their applications, even if it would be called something like *"The Encyclopedia of Sensors."* This is a different book and the author's task was much less ambitious. Here, an attempt has been made to generate a reference text, which could be used by students, researchers interested in modern instrumentation (applied physicists and engineers), sensor designers, application engineers and technicians whose job is to understand, select and/or design sensors for practical systems.

The prior editions of this book have been used quite extensively as desktop references and textbooks for the related college courses. Comments and suggestions from the sensor designers, professors, and students prompted me to implement several changes and correct errors. I am deeply grateful to those who helped me to make further improvements in this new edition. I owe a debt of gratitude and many thanks to Drs. Ephraim Suhir and David Pintsov for assisting me in mathematical treatment of transfer functions and to Drs. Todd E. Mlsna and Sanjay V. Patel for their invaluable contribution to the chapter on chemical sensors.

Even though the book is intended for the scientific and engineering communities, as a rule, technical descriptions and mathematic treatments do not require a background beyond a high school curriculum. Simplicity of description and intuitive approach were the key requirements that I set for myself while working on the manuscript. My true goal was not to pile up a collection of information but rather to entice the reader into a creative process. As Plutarch said nearly two millennia ago, "*The mind is not a vessel to be filled but a fire to be kindled. . .*"

San Diego, California Jacob Fraden
April, 2010

Contents

1 Data Acquisition 1
1.1 Sensors, Signals, and Systems 1
1.2 Sensor Classification 7
1.3 Units of Measurements 11
References 12

2 Sensor Characteristics 13
2.1 Transfer Function 13
2.1.1 Mathematical Model 14
2.1.2 Functional Approximations 15
2.1.3 Polynomial Approximations 16
2.1.4 Sensitivity 17
2.1.5 Linear Piecewise Approximation 18
2.1.6 Spline Interpolation 19
2.1.7 Multidimensional Transfer Functions 19
2.2 Calibration 20
2.2.1 Computation of Transfer Function Parameters 22
2.2.2 Linear Regression 25
2.3 Computation of Stimulus 26
2.3.1 Computation from Linear Piecewise Approximation 26
2.3.2 Iterative Computation of Stimulus (Newton Method) 28
2.4 Span (Full-Scale Full Scale Input) 30
2.5 Full-Scale Output 31
2.6 Accuracy 31
2.7 Calibration Error 34
2.8 Hysteresis 35
2.9 Nonlinearity 36
2.10 Saturation 37
2.11 Repeatability 38
2.12 Dead Band 38
2.13 Resolution 38

2.14 Special Properties 39
2.15 Output Impedance 40
2.16 Output Format 40
2.17 Excitation 41
2.18 Dynamic Characteristics 41
2.19 Environmental Factors 45
2.20 Reliability 47
2.21 Application Characteristics 49
2.22 Uncertainty 50
References 52

3 Physical Principles of Sensing 53
3.1 Electric Charges, Fields, and Potentials 54
3.2 Capacitance 60
3.2.1 Capacitor 62
3.2.2 Dielectric Constant 63
3.3 Magnetism 67
3.3.1 Faraday Law 69
3.3.2 Solenoid 71
3.3.3 Toroid 72
3.3.4 Permanent Magnets 72
3.4 Induction 73
3.5 Resistance 77
3.5.1 Specific Resistivity 79
3.5.2 Temperature Sensitivity 80
3.5.3 Strain Sensitivity 84
3.5.4 Moisture Sensitivity 85
3.6 Piezoelectric Effect 86
3.6.1 Ceramic Piezoelectric Materials 89
3.6.2 Polymer Piezoelectric Films 93
3.7 Pyroelectric Effect 96
3.8 Hall Effect 103
3.9 Thermoelectric Effects 106
3.9.1 Seebeck Effect 106
3.9.2 Peltier Effect 111
3.10 Sound Waves 113
3.11 Temperature and Thermal Properties of Materials 116
3.11.1 Temperature Scales 117
3.11.2 Thermal Expansion 118
3.11.3 Heat Capacity 120
3.12 Heat Transfer 121
3.12.1 Thermal Conduction 122
3.12.2 Thermal Convection 125
3.12.3 Thermal Radiation 126

3.13 Light 135
3.13.1 Light Polarization 136
3.13.2 Light Scattering 137
3.14 Dynamic Models of Sensor Elements 138
3.14.1 Mechanical Elements 139
3.14.2 Thermal Elements 141
3.14.3 Electrical Elements 142
3.14.4 Analogies 143
References 144

4 Optical Components of Sensors 147
4.1 Radiometry 149
4.2 Photometry 154
4.3 Windows 157
4.4 Mirrors 158
4.5 Lenses 161
4.6 Fresnel Lenses 163
4.7 Fiber Optics and Waveguides 165
4.8 Concentrators 169
4.9 Coatings for Thermal Absorption 170
4.10 Nano-optics 172
References 172

5 Interface Electronic Circuits 173
5.1 Input Characteristics of Interface Circuits 173
5.2 Amplifiers 178
5.2.1 Operational Amplifiers 178
5.2.2 Voltage Follower 181
5.2.3 Instrumentation Amplifier 182
5.2.4 Charge Amplifiers 183
5.3 Light-to-Voltage Converters 186
5.4 Excitation Circuits 188
5.4.1 Current Generators 188
5.4.2 Voltage References 192
5.4.3 Oscillators 192
5.4.4 Drivers 194
5.4.5 Optical Drivers 196
5.5 Analog-to-Digital Converters 196
5.5.1 Basic Concepts 196
5.5.2 V/F Converters 198
5.5.3 Dual-Slope Converters 203
5.5.4 Successive Approximation Converter 203
5.5.5 Resolution Extension 205
5.6 Direct Digitization 207

5.7 Capacitance-to-Voltage Converters 208
5.8 Integrated Interfaces 210
5.9 Ratiometric Circuits 211
5.10 Differential Circuits 214
5.11 Bridge Circuits 215
5.11.1 General Concept 215
5.11.2 Disbalanced Bridge 216
5.11.3 Null-Balanced Bridge 218
5.11.4 Bridge Amplifiers 218
5.12 Data Transmission 220
5.12.1 Two-Wire Transmission 220
5.12.2 Four-Wire Sensing 221
5.12.3 Six-Wire Sensing 222
5.13 Noise in Sensors and Circuits 223
5.13.1 Inherent Noise 223
5.13.2 Transmitted Noise 227
5.13.3 Electric Shielding 231
5.13.4 Bypass Capacitors 234
5.13.5 Magnetic Shielding 235
5.13.6 Mechanical Noise 237
5.13.7 Ground Planes 237
5.13.8 Ground Loops and Ground Isolation 238
5.13.9 Seebeck Noise 240
5.14 Calibration 242
5.15 Batteries for Low-Power Sensors 243
5.15.1 Primary Cells 244
5.15.2 Secondary Cells 245
References 246

6 Occupancy and Motion Detectors 247
6.1 Ultrasonic Detectors 249
6.2 Microwave Motion Detectors 249
6.3 Capacitive Occupancy Detectors 254
6.4 Triboelectric Detectors 258
6.5 Optoelectronic Motion Detectors 260
6.5.1 Sensor Structures 261
6.5.2 Visible and Near IR Light Motion Detectors 264
6.5.3 Far-Infrared Motion Detectors 267
6.6 Optical Presence Sensors 274
6.7 Pressure-Gradient Sensors 276
References 278

7 Position, Displacement, and Level 279
7.1 Potentiometric Sensors 280
7.2 Capacitive Sensors 284

7.3 Inductive and Magnetic Sensors 288
7.3.1 LVDT and RVDT 288
7.3.2 Eddy Current Sensors 290
7.3.3 Transverse Inductive Sensor 292
7.3.4 Hall Effect Sensors 293
7.3.5 Magnetoresistive Sensors 297
7.3.6 Magnetostrictive Detector 300
7.4 Optical Sensors 302
7.4.1 Optical Bridge 302
7.4.2 Proximity Detector with Polarized Light 303
7.4.3 Fiber-Optic Sensors 304
7.4.4 Fabry-Perot Sensors 306
7.4.5 Grating Sensors 308
7.4.6 Linear Optical Sensors 310
7.5 Ultrasonic Sensors 314
7.6 Radar Sensors 316
7.6.1 Micropower Impulse Radar 316
7.6.2 Ground Penetrating Radars 318
7.7 Thickness and Level Sensors 320
7.7.1 Ablation Sensors 320
7.7.2 Thin Film Sensors 322
7.7.3 Liquid Level Sensors 323
7.8 Pointing Devices 324
7.8.1 Optical Pointing Devices 324
7.8.2 Magnetic Pickup 325
7.8.3 Inertial and Gyroscopic Mice 325
References 325

8 Velocity and Acceleration 327
8.1 Accelerometer Characteristics 329
8.2 Capacitive Accelerometers 332
8.3 Piezoresistive Accelerometers 334
8.4 Piezoelectric Accelerometers 335
8.5 Thermal Accelerometers 336
8.5.1 Heated Plate Accelerometer 336
8.5.2 Heated Gas Accelerometer 337
8.6 Gyroscopes 339
8.6.1 Rotor Gyroscope 340
8.6.2 Monolithic Silicon Gyroscopes 341
8.6.3 Optical (Laser) Gyroscopes 344
8.7 Piezoelectric Cables 346
8.8 Gravitational Sensors 348
References 351

9 Force, Strain, and Tactile Sensors 353
9.1 Strain Gauges 355
9.2 Tactile Sensors 357
9.2.1 Switch Sensors 358
9.2.2 Piezoelectric Sensors 359
9.2.3 Piezoresistive Sensors 362
9.2.4 MEMS Sensors 364
9.2.5 Capacitive Touch Sensors 365
9.2.6 Acoustic Touch Sensors 369
9.2.7 Optical Sensors 369
9.3 Piezoelectric Force Sensors 370
References 372

10 Pressure Sensors 375
10.1 Concepts of Pressure 375
10.2 Units of Pressure 377
10.3 Mercury Pressure Sensor 378
10.4 Bellows, Membranes, and Thin plates 379
10.5 Piezoresistive Sensors 381
10.6 Capacitive Sensors 387
10.7 VRP Sensors 388
10.8 Optoelectronic Pressure Sensors 390
10.9 Indirect Pressure Sensor 391
10.10 Vacuum Sensors 393
10.10.1 Pirani Gauge 393
10.10.2 Ionization Gauges 395
10.10.3 Gas Drag Gauge 396
10.10.4 Membrane Vacuum Sensors 396
References 397

11 Flow Sensors 399
11.1 Basics of Flow Dynamics 399
11.2 Pressure Gradient Technique 402
11.3 Thermal Transport Sensors 404
11.3.1 Hot-Wire Anemometers 405
11.3.2 Three-Part Thermoanemometer 409
11.3.3 Two-Part Thermoanemometer 411
11.3.4 Microflow Thermal Transport Sensors 414
11.4 Ultrasonic Sensors 416
11.5 Electromagnetic Sensors 418
11.6 Breeze Sensor 420
11.7 Coriolis Mass Flow Sensors 422
11.8 Drag Force Sensors 423
11.9 Dust and Smoke Detectors 424

11.9.1 Ionization Detector 424
11.9.2 Optical Detector 426
References 428

12 Acoustic Sensors 431
12.1 Resistive Microphones 432
12.2 Condenser Microphones 432
12.3 Fiber-Optic Microphone 434
12.4 Piezoelectric Microphones 435
12.5 Electret Microphones 437
12.6 Dynamic Microphones 439
12.7 Solid-State Acoustic Detectors 440
References 443

13 Humidity and Moisture Sensors 445
13.1 Concept of Humidity 445
13.2 Capacitive Sensors 448
13.3 Electrical Conductivity Sensors 452
13.4 Thermal Conductivity Sensor 455
13.5 Optical Hygrometer 456
13.6 Oscillating Hygrometer 458
References 459

14 Light Detectors 461
14.1 Introduction 461
14.2 Photodiodes 465
14.3 Phototransistor 471
14.4 Photoresistors 472
14.5 Cooled Detectors 475
14.6 Image Sensors 478
14.6.1 CCD Sensor 479
14.6.2 CMOS-Imaging Sensors 480
14.7 Thermal Detectors 481
14.7.1 Golay Cells 482
14.7.2 Thermopile Sensors 483
14.7.3 Pyroelectric Sensors 487
14.7.4 Bolometers 491
14.7.5 Active Far-Infrared Sensors 494
14.8 Optical Design 497
14.9 Gas Flame Detectors 498
References 500

15 Radiation Detectors 503
15.1 Scintillating Detectors 504

15.2 Ionization Detectors 507
15.2.1 Ionization Chambers 508
15.2.2 Proportional Chambers 509
15.2.3 Geiger–Müller Counters 510
15.2.4 Semiconductor Detectors 512
15.3 Cloud and Bubble Chambers 516
References 518

16 Temperature Sensors 519
16.1 Coupling with Object 519
16.2 Temperature Reference Points 526
16.3 Thermoresistive Sensors 528
16.3.1 Resistance Temperature Detectors 528
16.3.2 *Silicon* Resistive PTC Sensors 529
16.3.3 Thermistors 532
16.4 Thermoelectric Contact Sensors 549
16.4.1 Thermoelectric Laws 550
16.4.2 Thermocouple Circuits 552
16.4.3 Thermocouple Assemblies 554
16.5 Semiconductor *pn*-Junction Sensors 556
16.6 Optical Temperature Sensors 560
16.6.1 Fluoroptic Sensors 561
16.6.2 Interferometric Sensors 562
16.6.3 Thermochromic Solution Sensor 563
16.7 Acoustic Temperature Sensor 564
16.8 Piezoelectric Temperature Sensors 565
References 566

17 Chemical Sensors 569
17.1 Overview 570
17.2 History 570
17.3 Chemical Sensor Characteristics 571
17.4 Classes of Chemical Sensors 572
17.4.1 Electrical and Electrochemical Transducers 572
17.4.2 Elastomer Chemiresistors 581
17.4.3 Photoionization Detector 585
17.4.4 Physical Transducers 586
17.4.5 Optical Transducers 595
17.5 Biochemical Sensors 597
17.5.1 Enzyme Sensors 597
17.6 Multisensor Arrays 598
17.7 Electronic Noses and Tongues 599
17.8 Specific Difficulties 602
References 603

18 Sensor Materials and Technologies 607
18.1 Materials 607
18.1.1 Silicon as Sensing Material 607
18.1.2 Plastics 611
18.1.3 Metals 615
18.1.4 Ceramics 617
18.1.5 Glasses 617
18.1.6 Optical Glasses 618
18.1.7 Nanomaterials 620
18.2 Surface Processing 621
18.2.1 Deposition of Thin and Thick Films 621
18.2.2 Spin Casting 621
18.2.3 Vacuum Deposition 622
18.2.4 Sputtering 623
18.2.5 Chemical Vapor Deposition 624
18.2.6 Electroplating 625
18.3 Microtechnology 626
18.3.1 Photolithography 627
18.3.2 Silicon Micromachining 628
References 635

Appendix 637

Index 653

Chapter 1
Data Acquisition

It's as large as life, and twice as natural.

– Lewis Carroll, *Through the Looking Glass*

1.1 Sensors, Signals, and Systems

A sensor is often defined as a "device that receives and responds to a signal or stimulus." This definition is broad. In fact, it is so broad that it covers almost everything from a human eye to a trigger in a pistol. Consider the level-control system shown in Fig. 1.1 [1]. The operator adjusts the level of fluid in the tank by manipulating its valve. Variations in the inlet flow rate, temperature changes (these would alter the fluid's viscosity and consequently the flow rate through the valve), and similar disturbances must be compensated for by the operator. Without control, the tank is likely to flood, or run dry. To act appropriately, the operator must obtain timely information about the level of fluid in the tank. In this example, the information is generated by the sensor, which consists of two main parts: the sight tube on the tank and the operator's eye, which produces an electric response in the optic nerve. The sight tube by itself is not a sensor, and in this particular control system, the eye is not a sensor either. Only the combination of these two components makes a narrow-purpose sensor (detector), which is *selectively* sensitive to the fluid level. If a sight tube is designed properly, it will very quickly reflect variations in the level, and it is said that the sensor has a fast speed response. If the internal diameter of the tube is too small for a given fluid viscosity, the level in the tube may lag behind the level in the tank. Then, we have to consider a phase characteristic of such a sensor. In some cases, the lag may be quite acceptable, while in other cases, a better sight tube design would be required. Hence, the sensor's performance must be assessed only as part of a data acquisition system.

This world is divided into natural and human-made objects. The natural sensors, like those found in living organisms, usually respond with signals, having an

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_1, © Springer Science+Business Media, LLC 2010

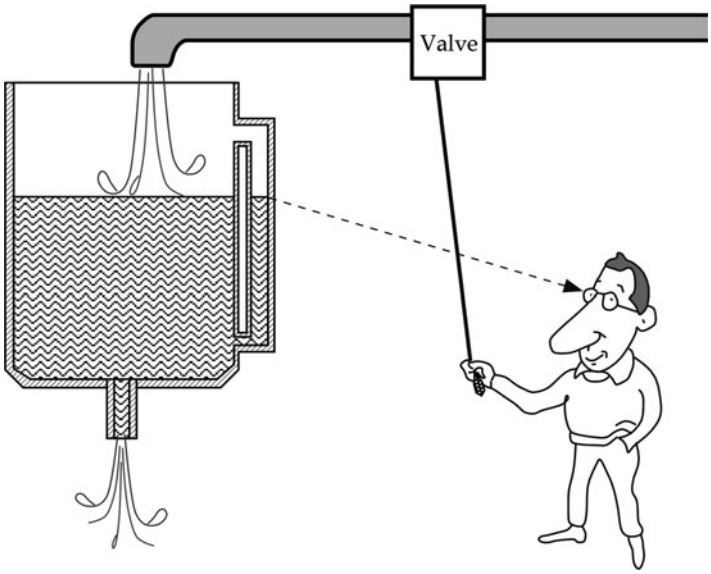

Fig. 1.1 Level control system. A sight tube and operator's eye form a sensor, a device which converts information into an electrical signal

electrochemical character, that is, their physical nature is based on ion transport, like in the nerve fibers (such as an optic nerve in the fluid tank operator). In man-made devices, information is also transmitted and processed in electrical form, however, through the transport of electrons. Sensors that are used in the artificial systems must speak the same language as the devices with which they are interfaced. This language is electrical in its nature and a man-made sensor should be capable of responding with signals where information is carried by displacement of electrons, rather than ions.[1] Thus, it should be possible to connect a sensor to an electronic system through electrical wires rather than through an electrochemical solution or a nerve fiber. Hence, in this book, we use a somewhat narrower definition of sensors, which may be phrased as

A sensor is a device that receives a stimulus and responds with an electrical signal.

The term *stimulus* is used throughout this book and needs to be clearly understood. The stimulus is the quantity, property, or condition that is received and converted into an electrical signal. Some texts (for instance, [2]) use a different term, *measurand* which has the same meaning, however with the stress on quantitative characteristic of sensing.

[1]There is a very exciting field of the optical computing and communications where information is processed by a transport of photons. That field is beyond the scope of this book.

The purpose of a sensor is to respond to some kind of an input physical property (stimulus) and to convert it into an electrical signal that is compatible with electronic circuits. We may say that a sensor is a translator of a generally nonelectrical value into an electrical value. When we say "electrical," we mean a signal, which can be channeled, amplified, and modified by electronic devices. The sensor's output signal may be in the form of voltage, current, or charge. These may be further described in terms of amplitude, polarity, frequency, phase, or digital code. This set of characteristics is called the *output signal format*. Therefore, a sensor has input properties (of any kind) and electrical output properties.

Any sensor is an energy converter. No matter what you try to measure, you always deal with energy transfer from the object of measurement to the sensor. The process of sensing is a particular case of information transfer, and any transmission of information requires transmission of energy. Of course, one should not be confused by an obvious fact that transmission of energy can flow both ways – it may be with a positive sign as well as with a negative sign; that is, energy can flow either from an object to the sensor or from the sensor to the object. A special case is when the net energy flow is zero, which also carries information about existence of that particular case. For example, a thermopile infrared radiation sensor will produce a positive voltage when the object is warmer than the sensor (infrared flux is flowing to the sensor) or the voltage is negative when the object is cooler than the sensor (infrared flux flows from the sensor to the object). When both the sensor and the object are at the same temperature, the flux is zero and the output voltage is zero. This carries a message that the temperatures are the same.

The term *sensor* should be distinguished from *transducer*. The latter is a converter of any one type of energy into another, whereas the former converts any type of energy into electrical energy. An example of a transducer is a loudspeaker, which converts an electrical signal into a variable magnetic field and, subsequently, into acoustic waves.[2] This is nothing to do with perception or sensing. Transducers may be used as *actuators* in various systems. An actuator may be described as an opposite to a sensor; it converts electrical signal into generally nonelectrical energy. For example, an electric motor is an actuator; it converts electric energy into mechanical action. Another example is a pneumatic actuator that is enabled by an electric signal.

Transducers may be parts of complex sensors (Fig. 1.2). For example, a chemical sensor may have a part, which converts the energy of a chemical reaction into heat (transducer) and another part, a thermopile, which converts heat into an electrical signal. The combination of the two makes a chemical sensor, a device which produces electrical signal in response to a chemical reagent. Note that in the above example a chemical sensor is a complex sensor; it is comprised of a nonelectrical transducer and a simple (direct) sensor converting heat to electricity. This suggests that many sensors incorporate at least one direct-type sensor and a

[2]It is interesting to note that a loudspeaker, when connected to an input of an amplifier, may function as a microphone. In that case, it becomes an acoustical sensor.

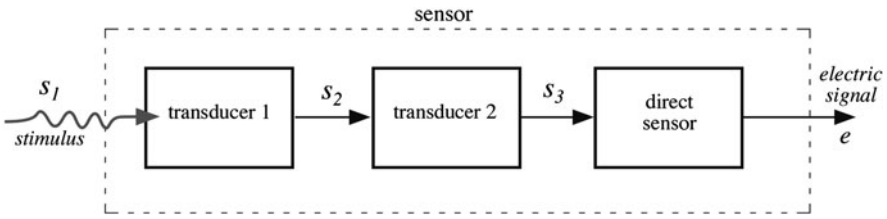

Fig. 1.2 A sensor may incorporate several transducers. s_1, s_2, and so on are various types of energy. Note that the last part is a direct sensor producing electrical output e

number of transducers. The direct sensors are those that employ certain physical effects to make a direct energy conversion into an electrical signal generation or modification. Examples of such physical effects are photoeffect and Seebeck effect. These will be described in Chap. 3.

In summary, there are two types of sensors; direct and complex. A direct sensor converts a stimulus into an electrical signal or modifies an electrical signal by using an appropriate physical effect, whereas a complex sensor in addition needs one or more transducers of energy before a direct sensor can be employed to generate an electrical output.

A sensor does not function by itself; it is always a part of a larger system that may incorporate many other detectors, signal conditioners, signal processors, memory devices, data recorders, and actuators. The sensor's place in a device is either intrinsic or extrinsic. It may be positioned at the input of a device to perceive the outside effects and to signal the system about variations in the outside stimuli. Also, it may be an internal part of a device that monitors the devices' own state to cause the appropriate performance. A sensor is always a part of some kind of a data acquisition system. Often, such a system may be a part of a larger control system that includes various feedback mechanisms.

To illustrate the place of sensors in a larger system, Fig. 1.3 shows a block diagram of a data acquisition and control device. An object can be anything: a car, space ship, animal or human, liquid, or gas. Any material object may become a subject of some kind of a measurement. Data are collected from an object by a number of sensors. Some of them (2, 3, and 4) are positioned directly on or inside the object. Sensor 1 perceives the object without a physical contact and, therefore, is called a noncontact sensor. Examples of such a sensor are a radiation detector and a TV camera. Even if we say "noncontact," we remember that energy transfer always occurs between any sensor and an object.

Sensor 5 serves a different purpose. It monitors internal conditions of a data acquisition system itself. Some sensors (1 and 3) cannot be directly connected to standard electronic circuits because of inappropriate output signal formats. They require the use of interface devices (signal conditioners). Sensors 1, 2, 3, and 5 are passive. They generate electric signals without energy consumption from the electronic circuits. Sensor 4 is active. It requires an operating signal, which is provided by an excitation circuit. This signal is modified by the sensor in

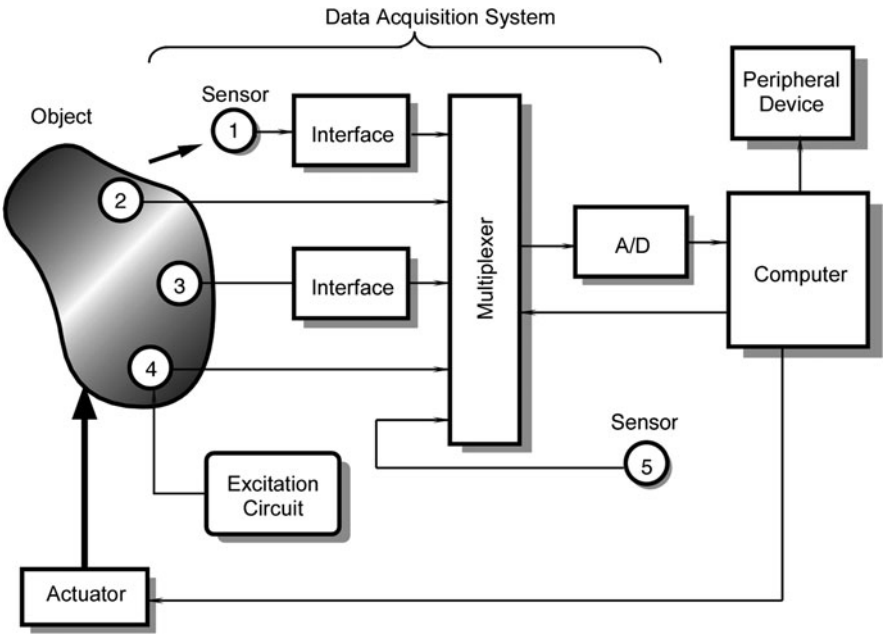

Fig. 1.3 Positions of sensors in a data acquisition system. Sensor 1 is noncontact, sensors 2 and 3 are passive, sensor 4 is active, and sensor 5 is internal to a data acquisition system

accordance with the converted information. An example of an active sensor is a thermistor, which is a temperature-sensitive resistor. It needs a constant current source, which is an excitation circuit. Depending on the complexity of the system, the total number of sensors may vary from as little as one (a home thermostat) to many thousands (a space shuttle).

Electrical signals from the sensors are fed into a multiplexer (MUX), which is a switch or a gate. Its function is to connect sensors one at a time to an analog-to-digital converter (A/D or ADC) if a sensor produces an analog signal, or directly to a computer if a sensor produces signals in a digital format. The computer controls a multiplexer and an A/D converter for the appropriate timing. Also, it may send control signals to the actuator, which acts on the object. Examples of the actuators are an electric motor, a solenoid, a relay, and a pneumatic valve. The system contains some peripheral devices (for instance, a data recorder, a display, an alarm, etc.) and a number of components that are not shown in the block diagram. These may be filters, sample-and-hold circuits, amplifiers, and so forth.

To illustrate how such a system works, let us consider a simple car door monitoring arrangement. Every door in a car is supplied with a sensor, which detects the door position (open or closed). In most cars, the sensor is a simple electric switch. Signals from all door sensors go to the car's internal processor (no need for an A/D converter as all door signals are in a digital format: ones or

zeros). The processor identifies which door is open and sends an indicating signal to the peripheral devices (a dashboard display and an audible alarm). A car driver (the actuator) gets the message and acts on the object (closes the door).

An example of a more complex device is an anesthetic vapor delivery system. It is intended to control the level of anesthetic drugs delivered to a patient by means of inhalation during surgical procedures. The system employs several active and passive sensors. The vapor concentration of anesthetic agents (such as halothane, isoflurane, or enflurane) is selectively monitored by an active piezoelectric sensor installed into a ventilation tube. Molecules of anesthetic vapors add mass to the oscillating crystal in the sensor and change its natural frequency, which is a measure of vapor concentration. Several other sensors monitor the concentration of CO_2 to distinguish exhale from inhale, and temperature and pressure, to compensate for additional variables. All of these data are multiplexed, digitized, and fed into the microprocessor, which calculates the actual vapor concentration. An anesthesiologist presets a desired delivery level and the processor adjusts the actuators (the valves) to maintain anesthetics at the correct concentration.

Another example of a complex combination of various sensors, actuators, and indicating signals is shown in Fig. 1.4. It is an advanced safety vehicle (ASV) that was developed by Nissan. The system is aimed at increasing safety of a car. Among many others, it includes a drowsiness warning system and drowsiness relieving system. This may include the eyeball movement sensor and the driver head inclination detector. The microwave, ultrasonic, and infrared range measuring sensors are incorporated into the emergency braking advanced advisory system to illuminate the break lamps even before the driver brakes hard in an emergency, thus advising the driver of a following vehicle to take evasive action. The obstacle warning system includes both the radar and infrared (IR) detectors. The adaptive cruise control system works like this: if the driver approaches too closely to a preceding vehicle, the speed is automatically reduced to maintain a suitable safety distance. The pedestrian monitoring system detects and alerts the driver to the

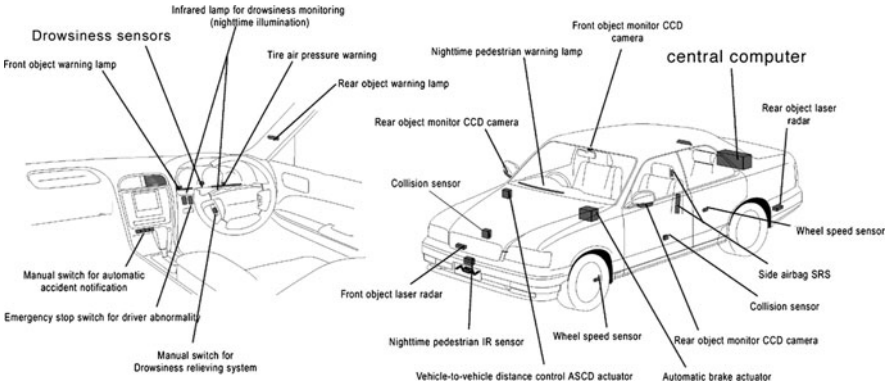
Drowsiness sensors
Infrared lamp for drowsiness monitoring (nighttime illumination)
Front object warning lamp
Tire air pressure warning
Rear object warning lamp
Manual switch for automatic accident notification
Emergency stop switch for driver abnormality
Manual switch for Drowsiness relieving system
Front object monitor CCD camera
Nighttime pedestrian warning lamp
central computer
Rear object laser radar
Rear object monitor CCD camera
Collision sensor
Wheel speed sensor
Side airbag SRS
Front object laser radar
Collision sensor
Nighttime pedestrian IR sensor
Wheel speed sensor
Rear object monitor CCD camera
Vehicle-to-vehicle distance control ASCD actuator
Automatic brake actuator

Fig. 1.4 Multiple sensors, actuators, and warning signals are parts of the advanced safety vehicle (Courtesy of Nissan Motor Company)

presence of pedestrians at night as well as in vehicle blind spots. The lane control system helps in the event that the system detects and determines that incipient lane deviation is not the driver's intention. It issues a warning and automatically steers the vehicle, if necessary, to prevent it from leaving its lane.

In the following chapters we concentrate on methods of sensing, physical principles of sensors operations, practical designs, and interface electronic circuits. Other essential parts of the control and monitoring systems, such as actuators, displays, data recorders, data transmitters, and others are beyond the scope of this book and mentioned only briefly.

The sensor's input signals (stimuli) may have almost any conceivable physical or chemical nature (e.g., light, temperature, pressure, vibration, displacement, position, velocity, ion concentration, etc). The sensor's design may be of a general purpose. A special packaging and housing should be built to adapt it for a particular application. For instance, a micromachined piezoresistive pressure sensor may be housed into a water-tight enclosure for the invasive measurement of aortic blood pressure through a catheter. The same sensor will be given an entirely different enclose when it is intended for measuring blood pressure by a noninvasive oscillometric method with an inflatable cuff. Some sensors are specifically designed to be very selective in a particular range of input stimulus and be quite immune to signals outside the desirable limits. For instance, a motion detector for a security system should be sensitive to movement of humans and not responsive to movement of smaller animals, like dogs and cats.

1.2 Sensor Classification

Sensor classification schemes range from very simple to the complex. Depending on the classification purpose, different classification criteria may be selected. Here, I offer several practical ways to look at the sensors.

1. All sensors may be of two kinds: passive and active. A passive sensor does not need any additional energy source and directly generates an electric signal in response to an external stimulus. That is, the input stimulus energy is converted by the sensor into the output signal. The examples are a thermocouple, a photodiode, and a piezoelectric sensor. Most of passive sensors are direct sensors as we defined them earlier.

The active sensors require external power for their operation, which is called an excitation signal. That signal is modified by the sensor to produce the output signal. The active sensors sometimes are called parametric because their own properties change in response to an external effect and these properties can be subsequently converted into electric signals. It can be stated that a sensor's parameter modulates the excitation signal and that modulation carries information of the measured value. For example, a thermistor is a temperature sensitive resistor. It does not generate

any electric signal, but by passing an electric current through it (excitation signal) its resistance can be measured by detecting variations in current and/or voltage across the thermistor. These variations (presented in ohms) directly relate to ttemperature through a known transfer function. Another example of an active sensor is a resistive strain gauge in which electrical resistance relates to a strain. To measure the resistance of a sensor, electric current must be applied to it from an external power source.

2. Depending on the selected reference, sensors can be classified into absolute and relative. An absolute sensor detects a stimulus in reference to an absolute physical scale that is independent of the measurement conditions, whereas a relative sensor produces a signal that relates to some special case. An example of an absolute sensor is a thermistor, a temperature-sensitive resistor. Its electrical resistance directly relates to the absolute temperature scale of Kelvin. Another very popular temperature sensor thermocouple is a relative sensor. It produces an electric voltage, which is a function of a temperature gradient across the thermocouple wires. Thus, a thermocouple output signal cannot be related to any particular temperature without referencing to a known baseline. Another example of the absolute and relative sensors is a pressure sensor. An absolute pressure sensor produces signal in reference to vacuum – an absolute zero on a pressure scale. A relative pressure sensor produces signal with respect to a selected baseline that is not zero pressure, for example, to the atmospheric pressure.
3. Another way to look at a sensor is to consider some of its properties that may be of a specific interest. Below are the lists of various sensor characteristics that may be considered (Tables 1.1–1.7).

Table 1.1 Sensor specifications

Sensitivity	Stimulus range (span)
Stability (short- and long-term)	Resolution
Accuracy	Selectivity
Speed of response	Environmental conditions
Overload characteristics	Linearity
Hysteresis	Dead band
Operating life	Output format
Cost, size, weight	Other

Table 1.2 Sensor material

Inorganic	Organic
Conductor	Insulator
Semiconductor	Liquid gas or plasma
Biological substance	Other

Table 1.3 Detection means used in sensors

Biological
Chemical
Electric, magnetic or electromagnetic wave
Heat, temperature
Mechanical displacement or wave
Radioactivity, radiation
Other

Table 1.4 Conversion phenomena

Physical	Thermoelectric
	Photoelectric
	Photomagnetic
	Magnetoelectric
	Electromagnetic
	Thermoelastic
	Electroelastic
	Thermomagnetic
	Thermooptic
	Photoelastic
	Other
Chemical	Chemical transformation
	Physical transformation
	Electrochemical process Spectroscopy
	Other
Biological	Biochemical transformation, Physical transformation
	Effect on test organism Spectroscopy
	Other

Table 1.5 Field of applications

Agriculture	Automotive
Civil engineering, construction	Domestic, appliances
Distribution, commerce, finance	Environment, meteorology, security
Energy, power	Information, telecommunication
Health, medicine	Marine
Manufacturing	Recreation, toys
Military	Space
Scientific measurement	Other
Transportation (excluding automotive)	

Table 1.6 Stimulus

Stimulus	
Acoustic	Wave amplitude, phase, polarization
	Spectrum
	Wave velocity
	Other
Biological	Biomass (types, concentration, states)
	Other
Chemical	Components (identities, concentration, states)
	Other
Electric	Charge, current
	Potential, voltage
	Electric field (amplitude, phase, polarization, spectrum)
	Conductivity
	Permittivity
	Other
Magnetic	Magnetic field (amplitude, phase, polarization, spectrum)
	Magnetic flux
	Permeability
	Other
Optical	Wave amplitude, phase, polarization, spectrum
	Wave velocity
	Refractive index
	Emissivity, reflectivity, absorption
	Other
Mechanical	Position (linear, angular)
	Acceleration
	Force
	Stress, pressure
	Strain
	Mass, density
	Moment, torque
	Speed of flow, rate of mass transport
	Shape, roughness, orientation
	Stiffness, compliance
	Viscosity
	Crystallinity, structural integrity
	Other
Radiation	Type
	Energy
	Intensity
	Other
Thermal	Temperature
	Flux
	Specific heat
	Thermal conductivity
	Other

Table 1.7 SI basic units

Quantity	Name	Symbol	Defined by... (year established)
Length	Meter	m	...the length of the path traveled by light in vacuum in 1/299,792,458 of a second... (1983)
Mass	Kilogram	kg	...after a platinum-iridium prototype (1889)
Time	Second	s	...the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (1967)
Electric current	Ampere	A	Force equal to 2×10^{-7} newton per meter of length exerted on two parallel conductors in vacuum when they carry the current (1946)
Thermodynamic temperature	Kelvin	K	The fraction 1/273.16 of the thermodynamic temperature of the triple point of water (1967)
Amount of substance	Mole	mol	...the amount of substance which contains as many elementary entities as there are atoms in 0.012 kg of carbon 12 (1971)
Luminous intensity	Candela	cd	...intensity in the perpendicular direction of a surface of 1/600,000 m^2 of a blackbody at temperature of freezing Pt under pressure of 101,325 newton per m^2 (1967)
Plane angle	Radian	rad	(supplemental unit)
Solid angle	Steradian	sr	(supplemental unit)

1.3 Units of Measurements

In this book, we use base units that have been established in The 14th General Conference on Weights and Measures (1971). The base measurement system is known as SI, which stands for Le Systéme International d'Unités in French (Table 1.7) [4]. All other physical quantities are derivatives of these base units.[3] Some of them are listed in Table A-3.

Often it is not convenient to use base or derivative units directly; in practice quantities may be either too large or too small. For convenience in the engineering work, multiples and submultiples of the units are generally employed. They can be obtained by multiplying a unit by a factor from Table A-2 When pronounced, in all cases the first syllable is accented. For example, 1 ampere (A) may be multiplied by factor of 10^{-3} to obtain a smaller unit 1 milliampere (mA), which is one thousandth of an ampere.

Sometimes, two other systems of units are used. They are the Gaussian System and the British System, which in the United States is modified as the U.S. Customary System. The United States is the only developed country where SI still is not in common use. However, with the increase of world integration, international cooperation gains a strong momentum. Hence, it is appears unavoidable that America will convert to SI in the future, though maybe not in our lifetime. Still, in

[3]The SI is often called the modernized metric system.

this book, we will generally use SI units, however, for the convenience of the reader, the U.S. customary system units will be used in places where U.S. manufacturers employ them for the sensor specifications. For the conversion to SI from other systems[4] the reader may use Tables A-4 of the Appendix. To make a conversion, a non-SI value should be multiplied by a number given in the table. For instance, to convert acceleration of 55 ft/s^2 to SI, it must to be multiplied by 0.3048:

$$55\,\text{ft/s}^2 \times 0.3048 = 16.764\ \text{m/s}^2$$

Similarly, to convert electric charge of 1.7 faraday, it must be multiplied by 9.65×10^{19}:

$$1.7\ \text{faraday} \times\ 9.65 \times\ 10^{19} = 1.64 \times\ 10^{20}\ \text{C}$$

The reader should consider the correct terminology of the physical and technical terms. For example, in the United States and many other countries, electric potential difference is called "voltage," while in other countries "electric tension" or simply "tension" is in common use (for example: *Spannung* in German and *напряжение* in Russian). In this book, we use terminology that is traditional in the United States of America.

References

1. Thompson S (1989) Control systems: engineering and design. Longman Scientific & Technical, Essex, England
2. Norton HN (1989) Handbook of transducers. Prentice Hall, Englewood Cliffs, NJ
3. White RW (1991) A sensor classification scheme. In: Microsensors. IEEE Press, New York, pp 3–5
4. Thompson A, Taylor BN (2008) Guide for the use of the international system of units (SI). NIST Special Publication 811, National Institute of Standards and Technology, Gaithersburg, MD 20899

[4]Nomenclature, abbreviations, and spelling in the conversion tables are in accordance with ASTM SI10–02 IEEE/ASTM SI10 American National Standard for Use of the International System of Units (SI): The Modern Metric System. A copy is available from ASTM, 100 Barr Harbor Dr., West Conshocken, PA 19428–2959, USA. Tel.: (610) 832–9585, www.astm.org/Standards/SI10.htm

Chapter 2
Sensor Characteristics

O, what men dare do! What men may do!
What men daily do, not knowing what they do.

- Shakespeare, *Much Ado About Nothing*

Since most of stimuli are not electrical, from an input to the output, a sensor may have several energy conversion steps before it produces and outputs an electrical signal. For example, pressure inflicted on a fiber optic pressure sensor, first results in strain in the fiber, which, in turn, causes deflection in its refractive index, which, in turn, results in an overall change in optical transmission and modulation of photon density. Finally, photon flux is detected by a photodiode and converted into electric current. In this chapter, we discuss the overall sensor characteristics, regardless of a physical nature or steps that are required to make energy conversions. Here, we consider a sensor as a "black box" where we concern only with relationships between its output electrical signal and input stimulus. Also, we will discuss the key point of sensing: computation of the input stimulus value from a measured sensor's electric output.

2.1 Transfer Function

An ideal or theoretical input–output (stimulus–response) relationship exists for every sensor. If a sensor is ideally designed and fabricated with ideal materials by ideal workers working in an ideal environment using ideal tools, the output of such a sensor would always represent the *true* value of the stimulus. This ideal input–output relationship may be expressed in the form of a table of values, a graph, a mathematical formula, or as a solution of a mathematical equation. If the input–output function is time invariant it is commonly called transfer function. This term is used throughout this book.

The transfer function represents the relation between stimulus s and response electrical signal S produced by the sensor. This relation can be written as $S = f(s)$.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_2, © Springer Science+Business Media, LLC 2010

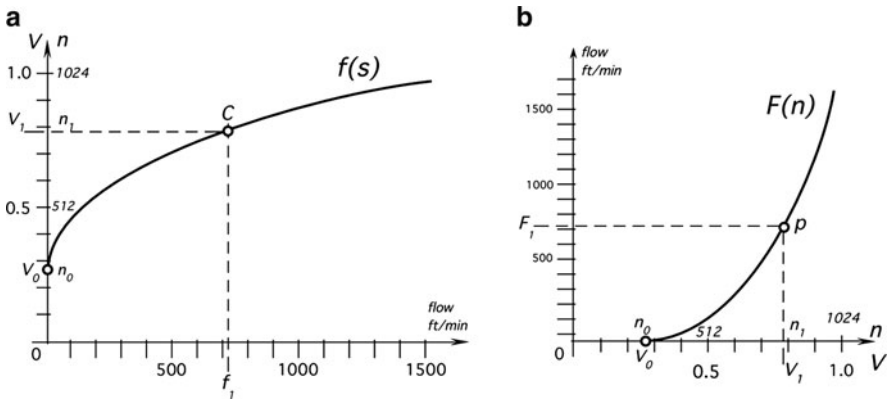

Fig. 2.1 Transfer function (**a**) and inverse transfer function (**b**) of a thermo-anemometer

Normally, stimulus s is unknown while the output signal S is measured. An inverse $f^{-1}(S)$ of the transfer function is required to compute the stimulus from the sensor's response S.

The value of S that becomes known during the measurement is just a number (voltage, current, digital count, etc.) that represents the value of stimulus s. In reality, any sensor is attached to a measuring system. One of the jobs of the system is to "break the code S" and infer the unknown value of s from the measured value of S. Thus in the measurement system an inverse transfer function $f^{-1}(S)$, which will be denoted $F(S)$, is employed to obtain the value of the stimulus s. Figure 2.1a illustrates the transfer function of a thermo-anemometer (a sensor that measures mass flow of gas). In general, it can be modeled by a square root function $f(s)$ of the input airflow rate. The output of the sensor can be in volts or in digital counts from an analog-to-digital (A/D) converter, as shown on the y-axis of Fig. 2.1a for a 10-bit A/D converter. After the output count $n = f(s)$ is measured, it must be translated back to the flow rate. The monotonic square root function $f(s)$ has parabola $F(n)$ as its inverse. This parabola is shown in Fig. 2.1b illustrating the relation between the output counts (or volts) and the input flow rate. Graphically, the inverse function can be obtained by mirror reflection with respect to the bisector of the right angle formed by x and y-axes.

2.1.1 Mathematical Model

Preferably, a physical or chemical law that forms a basis for the sensor's operation should be known. If such a law can be expressed in form of a mathematical formula, often it can used to calculate the sensor's inversed transfer function by inverting the

formula and computing the unknown value of s from the measured S. For example, if a linear resistive potentiometer is used for sensing displacement d, an Ohm's law can be applied to compute the transfer function as illustrated in Chap. 7 (7.1). In this case, the response S is the measured voltage v and the inverse transfer function $F(S)$ can be given as

$$d = \frac{v}{E}D, \tag{2.1}$$

where E is the reference voltage and D is the maximum displacement (full scale); both are constants. From this function we can compute displacement d from the measured voltage v.

In practice, readily solvable formulas for many transfer functions, especially for complex sensors, do not exist and one has to resort to various approximations of the direct and inverse transfer functions, which are the subjects of the next section.

2.1.2 Functional Approximations

If the approximating function is selected first, the act of approximation can be seen as a curve-fitting of experimentally observed values to the calculated values of the approximating function. The approximating function should be simple enough for ease of computation and inversion. Here are some most popular functions used for approximation of nonlinear transfer functions.

The simplest transfer function is linear. We represent it by the following equation:

$$S = A + Bs, \tag{2.2}$$

corresponding to the straight line with intercept A, that is, the output signal at zero input signal s=0, and slope B, which is sometimes called sensitivity (since the larger this coefficient the greater the influence of the stimulus). The output S is one of the characteristics of the output electric signal. It may be its amplitude, phase, frequency, pulse-width modulation (PWM) or a digital code, depending on the sensor properties, signal conditioning, and interface circuit. Note that (2.2) assumes that the transfer function passes, at least theoretically, through zero value of the input stimulus. In many cases, this is not the case and it may be desirable to reference the sensor not to zero but to some more practical input reference value s_0. If the sensor response S_0 is known for that input reference (from calibration, for example), (2.2) can be rewritten in form:

$$S = S_0 + B(s - s_0) \tag{2.2a}$$

Very few sensors are truly linear. At least a small nonlinearity is always present, especially for a broad input range of the stimuli. Thus, (2.2) and (2.2a) are just a linear approximation of a nonlinear sensor's response. In many cases, when nonlinearity cannot be ignored, the transfer function can be approximated by a multitude of linear mathematical functions that we will discuss below in greater detail.

Logarithmic function and the corresponding inverse function are respectively:

$$\begin{aligned} S &= A + B \ln s \\ s &= \mathrm{e}^{\frac{S-A}{B}} \end{aligned} \tag{2.3}$$

Exponential function and its inverse are given by :

$$\begin{aligned} S &= A\mathrm{e}^{ks} \\ s &= \frac{1}{k} \ln \frac{S}{A} \end{aligned} \tag{2.4}$$

Power function and its inverse can be expressed as

$$\begin{aligned} S &= A + Bs^k \\ s &= \sqrt[k]{\frac{S-A}{B}}, \end{aligned} \tag{2.5}$$

where A, B are parameters and k is the power factor.

All three of the above approximations possess a small number of parameters that must be determined during calibration (see below). This property makes them rather convenient, provided that they can really fit the response of a particular sensor. It is always useful to have as small a number of unknown parameters as possible, not the least for the sake of a lower cost of sensor calibration.

2.1.3 Polynomial Approximations

A sensor may have such a transfer function that none of the above functional approximations would fit sufficiently well. A sensor designer with a reasonably good mathematical background and physical intuition may utilize some other suitable functional approximations, but if none is found, several old and reliable techniques may come in handy. One is a polynomial approximation, that is, a power series. It should be noted that any continuous function can be approximated by a power series. For example, the exponential function of (2.4) can be approximately

calculated by a third order polynomial by dropping all the higher terms of its series expansion[1]:

$$S = Ae^{ks} \approx A\left(1 + ks + \frac{k^2}{2!}s^2 + \frac{k^3}{3!}s^3\right) \tag{2.6}$$

In many cases it is sufficient to investigate approximation of a sensor's response by the 2nd and 3rd degree polynomials that can be expressed as

$$\begin{aligned} S &= a_2 s^2 + b_2 s + c_2 \\ S &= a_3 s^3 + b_3 s^2 + c_3 s + d_3 \end{aligned} \tag{2.7}$$

Of course, it should be appreciated that the quadratic (2nd order) polynomial is a special case of the 3rd degree polynomial just as the 1st order (linear) polynomial (2.2) is a special case of the quadratic polynomial with $a_{2,3} = b_3 = 0$.

Obviously, the same technique can be applied to the inverse transfer function. Thus, it also can be approximated by a second or third degree

$$\begin{aligned} s &= A_2 S^2 + B_2 S + C_2 \\ s &= A_3 S^3 + B_3 S^2 + C_3 S + D_3 \end{aligned} \tag{2.8}$$

The coefficients A, B, and C can be converted into coefficients a, b, and c, but the analytical conversion is rather cumbersome and rarely used. Instead, depending in the need, usually either a direct or inversed transfer function is approximated, but not both.

In some cases, especially when a high accuracy is required, the higher order polynomials should be considered because the higher the order of a polynomial the better the fit. Still, even a 2nd order polynomial often may yield a fit with a sufficient accuracy when applied to a relatively narrow range of input stimuli.

2.1.4 Sensitivity

Recall that a coefficient B in (2.2) and (2.2a) is called *sensitivity*. For a nonlinear transfer function, sensitivity B is not a fixed number, as would be the case in a linear transfer function. A nonlinear transfer function exhibits different sensitivities at

[1]This third-order polynomial approximation yields good approximation only for $ks \ll 1$. In general, the error of a power series approximation is subject of a rather non-trivial mathematical analysis. Luckily, in most practical situations that analysis is rarely needed.

different points in intervals of stimuli. In case of nonlinear transfer functions, the sensitivity is defined as a first derivative of the transfer function:

$$b_i(s_i) = \frac{dS(s_i)}{ds} \approx \frac{\Delta S_i}{\Delta s_i}, \tag{2.9}$$

where, traditionally Δs_i is a small increment of the input stimulus and ΔS_i is the corresponding change in the output S of the transfer function.

2.1.5 Linear Piecewise Approximation

A linear piecewise approximation is a powerful method to employ in a computerized data acquisition system. The idea behind it is to break up a nonlinear transfer function of any shape into sections and consider each such section being linear as described by (2.2) or (2.2a). Curved segments between the sample points (knots) demarcating the sections are replaced with straight line segments, thus greatly simplifying the behavior of the function between the sample points. In other words, the knots are graphically connected by straight lines. This can also be seen as polygonal approximation of the original nonlinear function. Figure 2.2 illustrates the linear piecewise approximation of a non-linear function with the knots at input values s_0, s_1, s_2, s_3, s_4, and the corresponding output values n_0, n_1, n_2, n_3, n_4 (digital counts from the A/D converter).

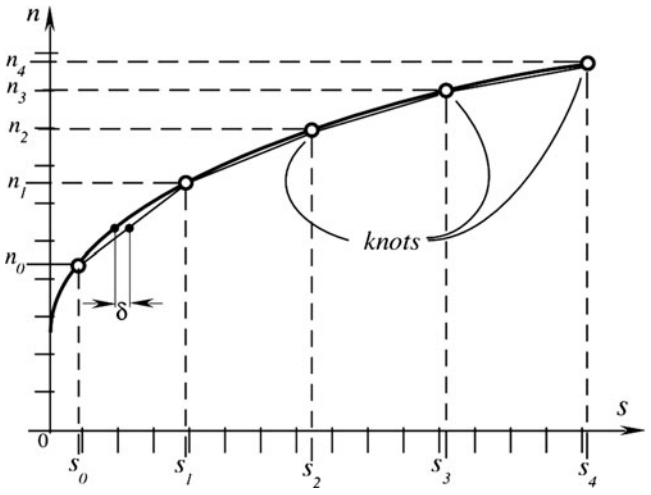

Fig. 2.2 Linear piecewise approximation

It makes sense to select knots only for the input range of interest (a *span* – see below), thus in Fig. 2.2 a section of the curve from 0 to s_0 is omitted as being outside of the required span limits.

An error of a piecewise approximation can be characterized by a maximum deviation δ of the approximation lines from the real curve. There exist different definitions of this maximum deviation (mean-square, absolute max, and others) but whatever the adopted metric, the larger δ calls for greater number of samples, that is larger number of sections with the idea of making this maximum deviation acceptably small. The knots do not need to be equally spaced. They should be closer to each other where a nonlinearity is high and farther apart where a nonlinearity is small.

2.1.6 Spline Interpolation

The approximation by higher order polynomials (3rd-order and higher) have some disadvantages: selected points at one side of the curve make strong influence on the remote parts of the curve. This deficiency is resolved by the spline method of approximation. In a similar way to a piecewise linear interpolation, the spline method is using a different 3rd order polynomial interpolation between the selected experimental points called knots. It is a curve between two neighboring knots and then all curves are "stitched" or "glued" together to obtain a smooth combined curve fitting. In fact, not necessarily it should be a 3rd-order curve – it can be as simple as the 1st order (linear) interpolation. A linear spline-interpolation (1st-order) is the simplest form and is equivalent to a linear piecewise interpolation as described above. The spline interpolation can utilize polynomials of different degrees, yet the most popular being cubic polynomials. Curvature of a line at each point is defined by the 2nd derivative. This derivative should be computed at each knot. If the 2nd derivatives are zero, the cubic spline is called "relaxed" and it is the choice for many practical approximations.

Spline interpolation is the efficient technique when it comes to interpolation that preserves smoothness of the transfer function. However, the simplicity of the implementation and the computational costs of spline interpolation should be taken into account particularly in a tightly controlled microprocessor environment.

2.1.7 Multidimensional Transfer Functions

A transfer function may be a function of more than one variable when the sensor's output is dependent on more than one input stimulus. One example is a humidity sensor whose output depends on two input variables – relative humidity and temperature. Another example is the transfer function of a thermal radiation (infrared)

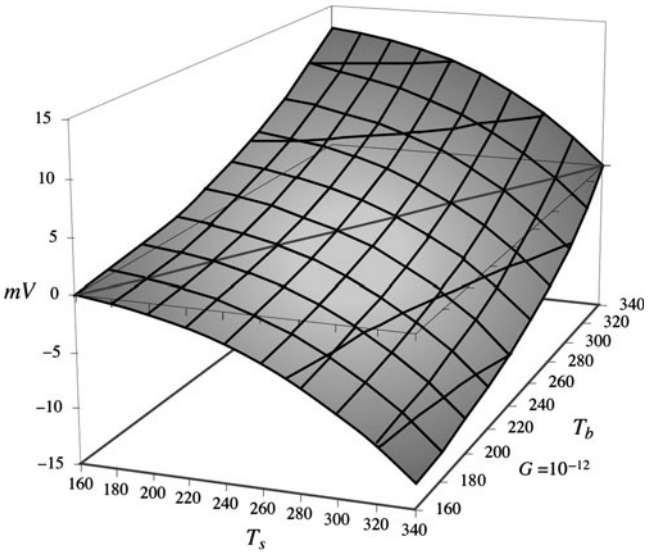

Fig. 2.3 Two-dimensional transfer function of a thermal radiation sensor

sensor. This function[2] has two arguments – two temperatures (T_b, the absolute temperature of an object of measurement and T_s, the absolute temperature of the sensor's surface), so the output voltage V is proportional to the difference

$$V = G\left(T_b^4 - T_s^4\right) \tag{2.10}$$

where G is a constant. Clearly, the relationship between the object's temperature and the output voltage (transfer function) is not only nonlinear (it depends on the 4th order parabola) but also depends on the sensor's surface temperature T_s, which should be measured by a separate temperature sensor. The graphical representation of a two-dimensional transfer function of 2.10 is shown in Fig. 2.3. It clearly depends on two input temperatures.

2.2 Calibration

If sensor's manufacturer tolerances and tolerances of the interface (signal conditioning) circuit are broader than the required system accuracy, a calibration of the sensor or a combination of a sensor and an interface circuit is required to minimize errors. For example, if one needs to measure temperature with accuracy ±0.1°C,

[2]This function is generally known as the Stefan-Boltzmann law (Sect. 3.12.3).

and the available sensor is rated as having accuracy of $\pm 1°C$ it does not mean that the sensor cannot be used. Rather this particular sensor needs calibration. That is, its unique transfer function should be found to fit the real sensor's response or the specific transfer function parameters should be adjusted to allow for a more accurate computation of the stimulus from the sensor's response.

There is no need to calibrate a sensor at many input stimuli. Usually, it is sufficient to calibrate only at a few sample points (stimuli) that are generated by a known reference source. The input and output points will lie on the real transfer function. The purpose of the calibration then is to find the unknown coefficients (parameters) of the inverted transfer function so that the fully defined function can be employed during the measurement process to compute any stimulus in the desirable range, not only at the points used during the calibration but anywhere in-between. In calibration, several input stimuli are paired with the corresponding output electric responses and the resulting pairs are plugged into the inverted transfer function to compute its parameters (coefficients). After the function parameters are established and stored, the sensor is ready for use.

Either a mathematical model of a transfer function has to be known before calibration or a good approximation of the sensor's response over the entire span must be found. In most cases, such functions are rather smooth and monotonic. Very rarely they contain singularities and if they do, such singularities are the useful phenomena that are employed for sensing (an ionizing particle detector is an example). Calibration of a sensor can be done in several possible ways, some of which are the following:

1. Calculation of the transfer function or its approximation to fit the selected calibration points (curve fitting by computing coefficients of a selected approximation).
2. Adjustment of the data acquisition system to trim (modify) the measured data by making them to fit into a normalized or "ideal" transfer function. An example is scaling of the acquired data.
3. Modification (trimming) of the sensor' properties to fit the predetermined transfer function.
4. Creating a sensor-specific reference device with matching properties at particular calibrating points.

Figure 2.4 illustrates three methods of calibrating a thermistor (temperature sensitive resistor).

In Fig. 2.4a, a thermistor is immersed into a stirred fluid bath with precisely controlled and monitored temperature. The fluid should be electrically nonconductive, such as oil or Fluorinert^{TM}. The temperature of the bath is monitored by a precision reference thermometer. The resistance of the thermistor is measured by an Ohm-meter, which is part of the calibration equipment. A grinder mechanically removes some material from the thermistor body to modify its dimensions and subsequently change its electrical resistance at the specific bath temperature. When the thermistor's resistance matches a predetermined value, the grinding stops and the calibration is finished. Now the thermistor response matches the "ideal" transfer

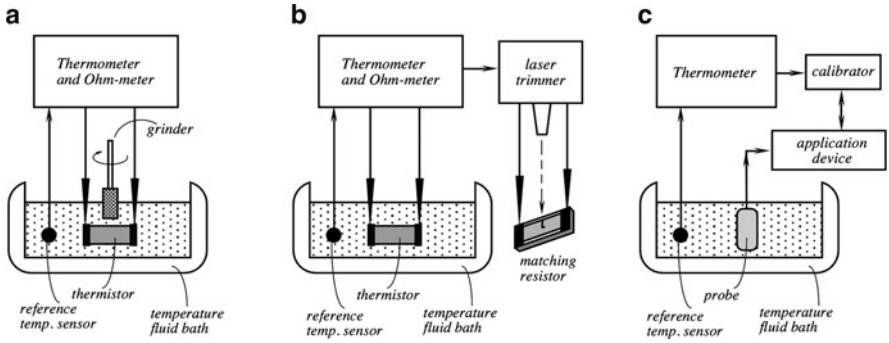

Fig. 2.4 Calibrations of a thermistor: grinding (**a**), trimming of a reference resistor (**b**), calculating the transfer function (**c**)

function. Another way of calibration is shown in Fig. 2.4b where the thermistor is not ground but just measured at a particular reference temperature. A regular temperature stable matching resistor is laser trimmed (or just selected from a stock) outside of the fluid bath. The trimming (selection) matches the resistor's resistance to that of the thermistor at a bath temperature. After the trimming, only that particular matched pair thermistor-resistor should be used in a measurement circuit, for example, in a Wheatstone bridge. Since it is a matching pair, the response of the bridge will correspond to an "ideal" transfer function.

In this example, the methods A and B are useful for calibration at one temperature point only, assuming that other parameters of the transfer function do not need calibration. If such is not the case, several calibrating pairs at different temperatures and resistances should be generated as shown in Fig. 2.4c. Here, the fluid bath is set at 2, 3, or 4 different temperatures and the sensor (imbedded into the probe) produces the corresponding responses, which are used by the calibrating device to generate the appropriate coefficients for the inverse transfer function which will be stored inside the application device (e.g., a thermometer).

2.2.1 Computation of Transfer Function Parameters

If a model of a transfer function is linear (2.2), then the calibration should determine constants A and B, if it is exponential (2.4), the constants A and k should be determined, and so on.

To calculate coefficients of the linear transfer function one needs two calibrating input–output pairs. Consider a simple linear transfer function of (2.2a). Since two points are required to define a straight line, a two-point calibration shall be performed. For example, if one uses a forward-biased semiconductor p–n junction

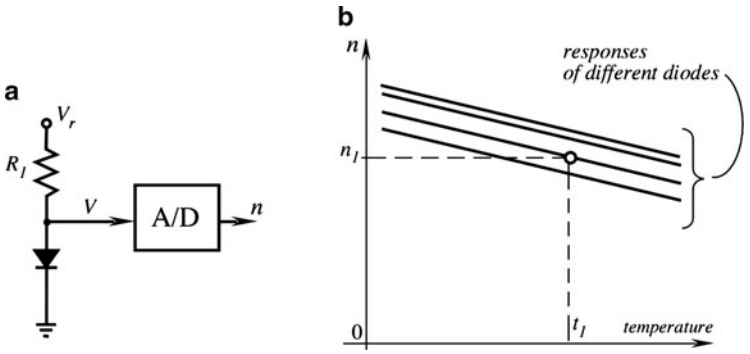

Fig. 2.5 A p–n junction temperature sensor (**a**); calibration (**b**). Each diode will produce different n_1 at the same temperature t_1. The slopes B are considered the same for all diodes

(Fig. 2.5a) as a temperature sensor (see Chap. 16), its transfer function is linear (temperature is the input and the A/D count n is the output):

$$n = n_1 + B(t - t_1) \tag{2.11}$$

The sensor shall be subjected to two calibrating temperatures (t_1 and t_2) and the two corresponding output counts (n_1 and n_2) will be registered. At first calibrating temperature t_1, the output count in n_1. After subjecting the sensor to the second calibrating temperature t_2, we arrive at

$$n_2 = n_1 + B(t_2 - t_1) \tag{2.12}$$

from which the sensitivity (slope) is computed as

$$B = \frac{n_2 - n_1}{t_2 - t_1} \tag{2.13}$$

and (2.11) becomes a linear transfer function with known parameters: B, n_1, and t_1.

Note that these parameters are unique for a particular sensor and must be stored in the measurement system. After calibration is done, temperature can be computed from the output counts n by use of the inversed transfer function

$$t = t_1 + \frac{(n - n_1)}{B} \tag{2.14}$$

In some fortunate cases, a constant B may be known upfront with a sufficient accuracy so that no computation of B is needed and one calibrating point is sufficient. In the same p–n junction of Fig. 2.5a, slope B has usually a very consistent value for a given lot and type of the semiconductor and can be considered

a known parameter. Yet, all diodes may have rather different offsets, so a single point calibration is needed to find out n_I at a calibrating temperature t_1.

For nonlinear transfer functions, calibration at one data point may be sufficient only in some rare cases, but often two and more input–output pairs would be required. When a 2nd or a 3rd degree polynomial transfer functions are employed, respectively 3 and 4 calibrating pairs are required. For a 3rd order polynomial

$$S = as^3 + bs^2 + cs + d \tag{2.15}$$

to find four parameters *a, b, c*, and *d*, four calibrating input–output pairs are required:

s_1 and S_1, s_2 and S_2, s_3 and S_3, s_4 and S_4.

Plugging these experimental pairs into (2.15) we get a system of four equations

$$\begin{aligned} S_1 &= as_1^3 + bs_1^2 + cs_1 + d \\ S_2 &= as_2^3 + bs_2^2 + cs_2 + d \\ S_3 &= as_3^3 + bs_3^2 + cs_3 + d \\ S_4 &= as_4^3 + bs_4^2 + cs_4 + d \end{aligned} \tag{2.16}$$

To solve this system for the parameters, first one computes the determinants of the systems:

$$\begin{aligned} \Delta &= \left(\frac{s_1^2 - s_2^2}{s_1 - s_2} - \frac{s_1^2 - s_4^2}{s_1 - s_4}\right)\left(\frac{s_1^3 - s_2^3}{s_1 - s_2} - \frac{s_1^3 - s_3^3}{s_1 - s_3}\right) \\ &\quad - \left(\frac{s_1^2 - s_2^2}{s_1 - s_2} - \frac{s_1^2 - s_3^2}{s_1 - s_3}\right)\left(\frac{s_1^3 - s_2^3}{s_1 - s_2} - \frac{s_1^3 - s_4^3}{s_1 - s_4}\right) \\ \Delta_a &= \left(\frac{s_1^2 - s_2^2}{s_1 - s_2} - \frac{s_1^2 - s_4^2}{s_1 - s_4}\right)\left(\frac{S_1 - S_2}{s_1 - s_2} - \frac{S_1 - S_3}{s_1 - s_3}\right) \\ &\quad - \left(\frac{s_1^2 - s_2^2}{s_1 - s_2} - \frac{s_1^2 - s_3^2}{s_1 - s_3}\right)\left(\frac{S_1 - S_2}{s_1 - s_2} - \frac{S_1 - S_4}{s_1 - s_4}\right) \\ \Delta_b &= \left(\frac{s_1^3 - s_2^3}{s_1 - s_2} - \frac{s_1^3 - s_3^3}{s_1 - s_3}\right)\left(\frac{S_1 - S_2}{s_1 - s_2} - \frac{S_1 - S_4}{s_1 - s_4}\right) \\ &\quad - \left(\frac{s_1^3 - s_2^3}{s_1 - s_2} - \frac{s_1^3 - s_4^3}{s_1 - s_4}\right)\left(\frac{S_1 - S_2}{s_1 - s_2} - \frac{S_1 - S_3}{s_1 - s_3}\right) \end{aligned} \tag{2.17}$$

from which the polynomial coefficients are calculated in the following fashion:

$$\begin{aligned} a &= \frac{\Delta_a}{\Delta}; \quad b = \frac{\Delta_b}{\Delta}; \\ c &= \frac{1}{s_1 - s_4}\left[S_1 - S_4 - a\left(s_1^3 - s_4^3\right) - b\left(s_1^2 - s_4^2\right)\right]; \\ d &= S_1 - as_1^3 - bs_1^2 - cs_1 \end{aligned} \tag{2.18}$$

If the determinant of the system Δ is small some considerable inaccuracy will result.

Since calibration may be a slow process (especially when dealing with either a large inertia or temperature), to reduce the manufacturing cost, it is important to minimize the number of calibration points. Thus, the most economical transfer function or the approximation should be selected. Economical means having the smallest number of unknown parameters. For example, if an acceptable accuracy can be achieved by a 2nd order polynomial, the 3rd order should not be used.

To calibrate sensors, it is essential to have and properly maintain precision and accurate references – physical standards of the appropriate stimuli. These references are the most critical parts of calibration equipment. For example, to calibrate contact temperature sensors either a temperature controlled fluid bath or a "dry well" cavity is required. For calibration of infrared radiation sensors, a blackbody cavity would be needed. In all cases, the calibration equipment must have a reference temperature sensor of the highest accuracy possible, preferably traceable to the national standard of temperature. To calibrate a hygrometer, a series of saturated salt solution is required to sustain a constant relative humidity in a closed container. It should be clearly understood that the calibration accuracy is directly linked to the accuracy of a reference sensor that is part of the calibration equipment. A value of uncertainty of the reference sensor should be included in the statement of the overall uncertainty, as explained below.

2.2.2 Linear Regression

If measurements of the input stimuli during calibration cannot be made with high accuracy and large random errors are expected, the minimal number of measurements will not yield a sufficient accuracy. For example, a two-point calibration for a linear transfer function would result in unacceptably high uncertainty. To cope with random errors in calibration process a method of *least squares* could be employed. Since this method is described in many textbooks and manuals, only the final expressions for the unknown parameters of a linear regression are given here for reminder. The reader is referred to any textbook on statistical error analysis.

Measure multiple (k) output values S at input values s over a substantially broad range, preferably over the entire span. Use the following formulas for a linear regression to determine intercept A and slope B of the best fitting straight line:

$$A = \frac{\Sigma S \Sigma s^2 - \Sigma s \Sigma s S}{k\Sigma s^2 - (\Sigma s)^2}, \quad B = \frac{k\Sigma s S - \Sigma s \Sigma S}{k\Sigma s^2 - (\Sigma s)^2}, \tag{2.19}$$

where Σ is the summation over all k pairs.

2.3 Computation of Stimulus

A goal of sensing is to determine a value of the input stimulus s from the value of the sensor output signal S. This can be done by two methods.

1. From the inverted transfer function $s = F(S)$ or its approximation, or
2. From a direct transfer function $S = f(s)$ by use of the iterative computation.

2.3.1 Computation from Linear Piecewise Approximation

During computation of stimulus s, the data acquisition system first determines between which knots the output signal lays and then treats the transfer function between the knots as a linear function with the known slope and intercept. Consider a linear approximation between knots p_1 and p_4 in Fig. 2.6.

A large triangle is formed with the corners at points p_1, p_2, and p_4. The unknown stimulus s_x corresponds to measured counts n_x. These values correspond to point p_5 on the approximation line, thus forming a smaller triangle between points p_1, p_2, and p_5. Both triangles are similar, which allows us to derive a linear equation for computing the unknown stimulus s_x from the measured value of n_x:

$$s_x = s_i + \frac{n_x - n_i}{n_{i+1} - n_i}(s_{i+1} - s_i) \tag{2.20}$$

This equation is easy to program and compute by an inexpensive microprocessor, which keeps in memory a look-up table containing the knots (Table 2.1).

For the illustration purpose, let us evaluate what would involve using a full functional model of a transfer function in comparison with a linear piecewise approximation. Figure 2.7a shows a thermistor temperature sensor with a pull-up resistor R_1 connected to a 12-bit analog-to-digital (A/D) converter (a full scale $N_0 =$

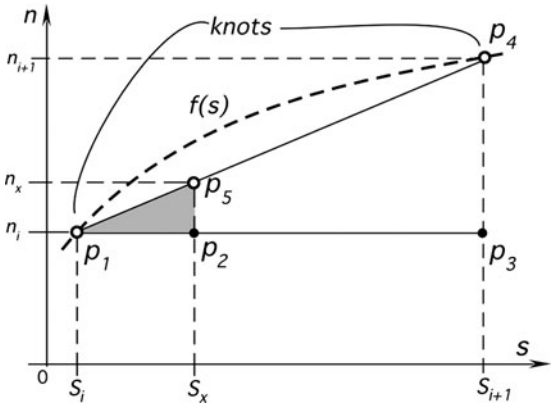

Fig. 2.6 Computation of stimulus from linear piecewise approximation

Table 2.1 Look-up table of knots for computing the input from the measured output

Knot	0	1	2	...	i	...	k
Output	n_0	n_1	n_2	...	n_i	...	n_k
Input	s_0	s_1	s_2	...	s_i	...	s_k

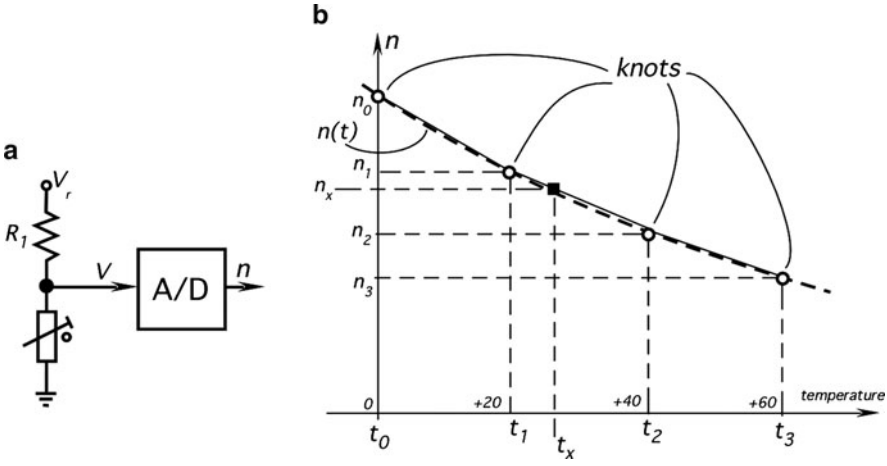

Fig. 2.7 Thermistor temperature sensor (**a**) and its linear piecewise approximation (**b**) with four knots

4,095 counts). We assume that the thermistor is used to measure temperature from 0°C to +60°C.

The output count from the A/D converter can be modeled by a nonlinear function $n(T)$ of temperature:

$$n_x = N_0 \frac{R_0 e^{\beta(T^{-1} - T_0^{-1})}}{R_1 + R_0 e^{\beta(T^{-1} - T_0^{-1})}}, \tag{2.21}$$

where T is the measured temperature in K, T_0 is the reference temperature in K, R_0 is the resistance of the thermistor at T_0, and in K is the characteristic temperature. After manipulating (2.21), we arrive at the inverted transfer function which enables us to compute analytically the input temperature in °K:

$$T_x = \left(\frac{1}{T_0} + \frac{1}{\beta} \ln\left(\frac{n_x}{N_0 - n_x} \frac{R_1}{R_0} \right) \right)^{-1} \tag{2.22}$$

First, we need to calibrate the sensor at two temperatures $T_x = T_{c1}$ and $T_x = T_{c2}$ in order to find out values of the constants R_0 and β. Assume that a pull-up resistor $= 10.00\ \text{k}\Omega$ is selected during design. Let us select two calibrating temperatures in the middle of the operating range as $T_0 = T_{c1} = 293.15$ K and $T_{c2} = 313.15$ K, which correspond to 20°C and 40°C. The thermistor sequentially is immersed into a liquid

Table 2.2 Look-up table of knots for computation of temperature

Knot	0	1	2	3
Counts	2,819	1,863	1,078	593
Temperature (°C)	0	20	40	60

bath at these two temperatures and the A/D counts are registered respectively as $n_{c1} = 1{,}863$ and $n_{c2} = 1{,}078$. By substituting these pairs into (2.21) and solving a system of two equations, we arrive at the coefficient values $R_0 = 8.350$ kΩ and $\beta = 3{,}895$ K. Now, the sensor is calibrated and (2.21) and (2.22) are fully characterized.

Since the sensor (thermistor) is calibrated and its specific transfer function is known, during the sensor use, (2.22) can be used for computing temperature from any reasonable A/D count. Yet (2.22) is not a simple equation to resolve by a microprocessor. To simplify computation, we may use a piecewise approximation. Let us break up the transfer function of (2.21) into three sections (Fig. 2.7b) with two end knots at 0°C and 60°C and two central knots at 20°C and 40°C. We will use linear approximations between the neighboring knot temperatures $t_0 = 0$°C and $t_1 = 20$°C, $t_1 = 20$°C and $t_2 = 40$°C, $t_2 = 40$°C and $t_3 = 60$°C. During the calibration, we compute the corresponding knot A/D counts $n_0 = 2{,}819$, $n_1 = 1{,}863$, $n_2 = 1{,}078$, and $n_3 = 593$. Alternatively, these values can be found from calibration at four points. The count–temperature pairs are plugged into a look-up Table 2.2.

As a practical example, consider that we use a thermistor to measure some unknown temperature and receive counts $n_x = 1{,}505$. From Table 2.2 we determine that this measured count is situated between knots 1 and 2. The find the temperature t_s, the measured counts and the knot values are plugged into (2.20) to arrive at

$$t_x = t_1 + \frac{n_x - n_1}{n_2 - n_i}(t_2 - t_1) = 20 + \frac{1505 - 1863}{1078 - 1863}(40 - 20) = 29.12 \qquad (2.23)$$

To check how far this computed temperature 29.12°C deviates from that computed from a "true" temperature, we plug the same $n_x = 1{,}505$ into (2.22) and compute $t_x = 28.22$°C. Hence, the selected linear piecewise approximation with two central knots overestimates temperature by 0.90°C. For some demanding applications, this may be a too much of error and thus more than two central knots would be required to bring the approximation error down to an acceptable level.

2.3.2 *Iterative Computation of Stimulus (Newton Method)*

Since the transfer function $S = f(s)$ can be rewritten as $S - f(s) = 0$, numerical iterative methods for finding roots of this typically nonlinear equation (such as Newton or secant methods [1, 2][3]) can also be used for calculating the unknown stimulus s without the knowledge of the inverse transfer function.

[3]This method is also known as the Newton–Raphson method, named after Isaac Newton and Joseph Raphson.

The method is based on first guessing an initial reasonable value of $s = s_0$ and then applying a Newton algorithm to compute a series of new values of s converging to the sought stimulus value. When a difference between two consecutively computed values of s becomes sufficiently small, the algorithm stops and the last computed value of s is considered a solution of the original equation and thus the value of the unknown stimulus is found. Newton's method converges remarkably quickly, especially if initial guess is reasonably close to the actual value of s.

If a sensor transfer function is $f(s)$, the Newton method prescribes computing for any measured output value S the following sequence of the stimuli values

$$s_{i+1} = s_i - \frac{f(s_i) - S}{f'(s_i)}, \tag{2.24}$$

which after several steps converges to the sought input s. Here s_{i+1} is the value at the iteration $i+1$, s_i is the value at a prior iteration i and $f'(s_i)$ is the first derivative of the transfer function at input s_i, $i = 0, 1, 2, \ldots$.

Use of (2.24) begins by guessing stimulus s_0. Then, computation of the subsequent s_i is performed several times (iterations) until the incremental change in s_i becomes sufficiently small, preferably in the range of the sensor resolution.

To illustrate the Newton method for a 3rd degree polynomial

$$f(s) = as^3 + bs^2 + cs + d \tag{2.25}$$

with coefficients $a = 1.5$, $b = 5$, $c = 25$, $d = 1$ (see Fig. 2.8) we plug (2.25) into (2.24) to arrive at an iteration of s_{i+1}:

$$s_{i+1} = s_i - \frac{as_i^3 + bs_i^2 + cs_i + d - S}{3as_i^2 + 2bs_i + c} = \frac{2as_i^3 + bs_i^2 - d + S}{3as_i^2 + 2bs_i + c}. \tag{2.26}$$

If, for instance, the measured sensor's response is $S = 22$ and our initial guess of the stimulus is $s_0 = 2$, then (2.26) will result in the following sequence of s_{i+1}:

$$\begin{aligned}
s_1 &= \frac{2 \cdot 1.5 \cdot 2^3 + 5 \cdot 2^2 - 1 + 22}{3 \cdot 1.5 \cdot 2^2 + 2 \cdot 5 \cdot 2 + 25} = 1.032 \\
s_2 &= \frac{2 \cdot 1.5 \cdot 1.032^3 + 5 \cdot 1.032^2 - 1 + 22}{3 \cdot 1.5 \cdot 1.032^2 + 2 \cdot 5 \cdot 1.032 + 25} = 0.738 \\
s_3 &= \frac{2 \cdot 1.5 \cdot 0.738^3 + 5 \cdot 0.738^2 - 1 + 22}{3 \cdot 1.5 \cdot 0.738^2 + 2 \cdot 5 \cdot 0.738 + 25} = 0.716 \\
s_4 &= \frac{2 \cdot 1.5 \cdot 0.716^3 + 5 \cdot 0.716^2 - 1 + 22}{3 \cdot 1.5 \cdot 0.716^2 + 2 \cdot 5 \cdot 0.716 + 25} = 0.716
\end{aligned} \tag{2.27}$$

We see that after just the third iteration, the sequence of s_i converges to 0.716. Hence, at step 4, the Newton algorithm stops and the stimulus value is deemed to

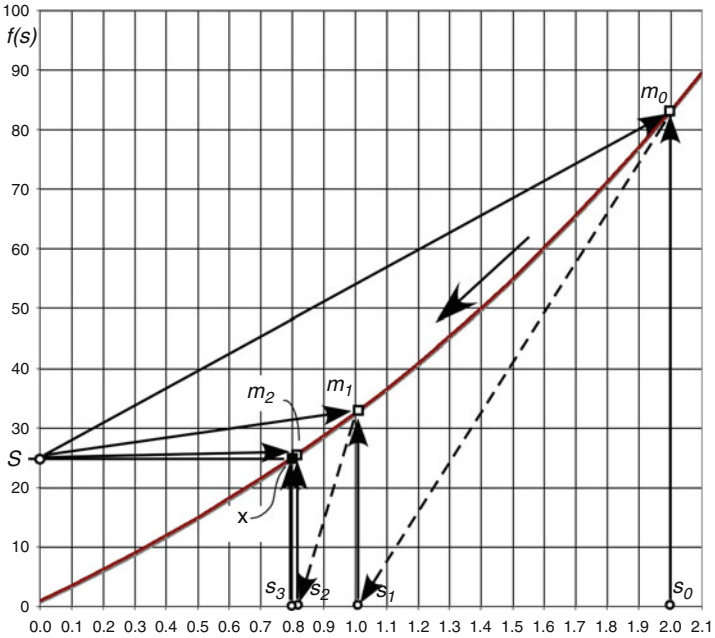

Fig. 2.8 Use of Newton method for computing sensor's stimulus. Computations starts from entering S and s_0 into (2.24) to compute stimulus s_1, which together with S generates s_2, and so on, till point x is computed

be $s = 0.716$. To check accuracy of this solution, plug this number into (2.25) and obtain $f(s) = S = 22.014$, which is within the resolution error (0.06%) of the actually measured response $S = 22$.

It should be noted that the Newton method results in large errors when the sensor's sensitivity becomes low. In other words, the method will fail where the transfer function flattens (1st derivative approaches zero). In such cases, the so-called Modified Newton Method may be employed. As was noted above, in some cases when the first derivative cannot be easily computed analytically, one uses instead a sensitivity value devised from Δs and ΔS (see (2.9)).

2.4 Span (Full-Scale Full Scale Input)

A dynamic range of stimuli that may be converted by a sensor is called a *span* or an *input full scale* (FS). It represents the highest possible input value, which can be applied to the sensor without causing unacceptably large inaccuracy. For the sensors with a very broad and nonlinear response characteristic, a dynamic range of the input stimuli is often expressed in decibels, which is a logarithmic measure of

Table 2.3 Relationship between power, force (voltage, current), and decibels

Power ratio	1.023	1.26	10.0	100	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
Force ratio	1.012	1.12	3.16	10.0	31.6	100	316	10^3	3162	10^4	3.10^4	10^5
Decibels	0.1	1.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0

ratios of either power or force (voltage). It should be emphasized that decibels do not measure absolute values, but a ratio of values only. A decibel scale represents signal magnitudes by much smaller numbers, which in many cases is far more convenient. Being a nonlinear scale, it may represent low level signals with high resolution while compressing the high level numbers. In other words, the logarithmic scale for small objects works as a microscope and for the large objects as a telescope. By definition, decibels are equal to ten times the log of the ratio of powers (Table 2.3):

$$1\ \text{dB} = 10 \log \frac{P_2}{P_1}. \tag{2.28}$$

In a similar manner, decibels are equal to 20 times the log of the force, or current, or voltage:

$$1\ \text{dB} = 20 \log \frac{S_2}{S_1} \tag{2.29}$$

2.5 Full-Scale Output

Full-scale output (FSO) is the algebraic difference between the electrical output signals measured with maximum input stimulus and the lowest input stimulus applied. This must include all deviations from the ideal transfer function. For instance, the FSO output in Fig. 2.9a is represented by *S*FS.

2.6 Accuracy

A very important characteristic of a sensor is accuracy, which really means inaccuracy. Inaccuracy is measured as a highest deviation of a value represented by the sensor from the ideal or true value of a stimulus at its input. The true value is attributed to the object of measurement and accepted as having a specified uncertainty (see below) because one never can be absolutely sure what the true value is.

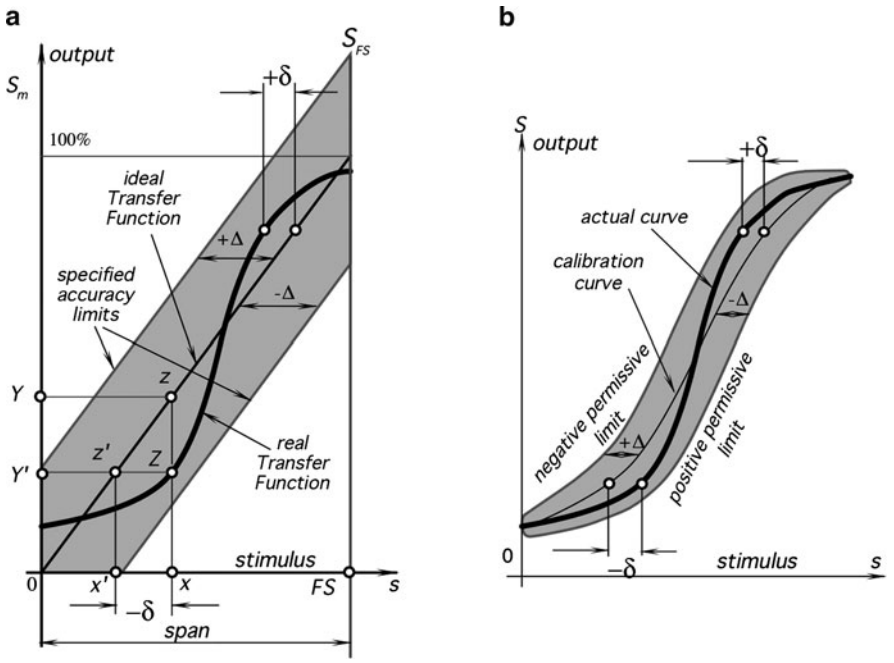

Fig. 2.9 Transfer function (**a**) and accuracy limits (**b**). Error is specified in terms of input values

The deviation can be described as a difference between the value, which is computed from the output voltage, and the actual input value. For example, a linear displacement sensor ideally should generate 1 mV per 1 mm displacement. That is, its transfer function is a line with a slope (sensitivity) $B = 1$ mV/mm. However, in the experiment, a reference displacement of $s = 10$ mm produced an output of $S = 10.5$ mV. Converting this number back into the displacement value by using the inverted transfer function ($1/B = 1$ mm/mV), we calculate the displacement as $s_x = S/B = 10.5$ mm. The result overestimates the displacement by $s_x - s = 0.5$ mm. This extra 0.5 mm is an erroneous deviation in the measurement, or error. Therefore, in a 10 mm range the sensor's absolute inaccuracy is 0.5 mm, or in relative terms the inaccuracy is 0.5 mm/10 mm times 100% = 5%. For a larger displacement, the error may be larger. If we repeat this experiment over and over again without any random error and every time we observe an error of 0.5 mm we may say that the sensor has a systematic inaccuracy of 0.5 mm over a 10 mm span. Naturally, a random component is always present, so the systematic error may be represented as an average or mean value of multiple errors.

Figure 2.9a shows an ideal or theoretical transfer function. In the real world, any sensor performs with some kind of imperfection. A possible real transfer function is represented by a thick line, which generally may be neither linear nor monotonic. A real function rarely coincides with the ideal. Because of the material variations,

workmanship, design errors, manufacturing tolerances, and other limitations, it is possible to have a large family of real transfer functions, even when sensors are tested under presumably identical conditions. However, all runs of the real transfer functions must fall within the limits of a specified accuracy. These permissive limits differ from the ideal transfer function line by $\pm\Delta$. The real functions deviate from the ideal by $\pm\delta$, where $\delta \leq \Delta$.

For example, let us consider a stimulus having value, x. Ideally, we would expect this value to correspond to point z on the transfer function, resulting in the output value Y. Instead, the real function will respond at point Z producing output value Y'. When we compute the value of a stimulus from the measured Y', we have no idea how the real transfer function differs from the expected "ideal" so we use the ideal inverted transfer function for the calculation. The measured output value Y' corresponds to point z' on the ideal transfer function, which, in turn, relates to a "would-be" input stimulus x' whose value is smaller than x. Thus, in this example imperfection in the sensor's performance leads to a measurement error $-\delta$.

The accuracy rating includes a combined effect of part-to-part variations, hysteresis, dead band, calibration, and repeatability errors (see below). The specified accuracy limits generally are used in the worst-case analysis to determine the worst possible performance of the system.

To improve accuracy, the number of the error-contributing factors should be reduced. This can be achieved by not relaying on the manufacturer's tolerances, but calibrating each sensor individually under selected conditions. Figure 2.9b shows that $\pm\Delta$ may more closely follow the real transfer function, meaning better sensor's accuracy. This can be accomplished by a multiple-point calibration of each individual sensor and curve fitting as described above. Thus, the specified accuracy limits are established not around the theoretical (ideal) transfer function, but around the actual calibration curve, which is adjusted during the calibration procedure. Then, the permissive limits become narrower as they do not embrace part-to-part variations between the sensors and are geared specifically to the particular device. Clearly, this method allows for a more accurate sensing, however in some applications, it may be prohibitive because of a higher cost.

Inaccuracy rating may be represented in a number of forms:

1. Directly in terms of measured value (Δ)

This form is used when error is independent on the input signal magnitude. Often, it relates to an additive noise or systematic bias, but also combines all other conceivable error sources, like calibration, manufacturer's tolerances, etc. For example, it can be stated as 0.15°C for a temperature sensor or 10 fpm (foot-per-minute) for a flow sensor.

2. In % of the input span (full scale)

This form is useful for a sensor with a linear transfer function and closely relates to the above form 1. It is just another way of stating the same thing because the span must be specified for nearly any sensor. This form is not useful for a sensor with

a nonlinear transfer function. For example, a thermoanemometer (see Sect. 11.3) has a response that can be modeled by a square root function, that is, it is more sensitive at low flow rates and less sensitive at high flows. Let us assume that the sensor[4] has span of 3000 fpm and its accuracy is stated as 3% of the full scale, which is the other way to say 90 fpm (see form 1 above). However, for measuring low flow rates, say from 30 to 100 fpm, this error of 90 fpm looks huge and in fact misleading due to nonlinearity.

3. In % of the measured signal

This is a multiplicative way of expressing error because the error magnitude is shown as fraction of the signal magnitude. It is useful for a sensor with a highly nonlinear transfer function. Considering the same example from the form 2 above, 3% of the measured signal is more practical for low-flow rates because it will be just few fpm, while for the high-flow rate it will be in tens of fpm, also reasonable. Still, using this form is not generally recommended. It makes more sense to break up the total nonlinear span into smaller quasilinear sections. Then, form 2 should be used instead for each individual section.

4. In terms of the output signal. This is useful for sensors with a digital output format so the error can be expressed, for example, in units of LSB

Which particular method to use often depends on the application.

In modern sensors, specification of accuracy often is replaced by a more comprehensive value of uncertainty (see Sect. 2.22) because uncertainty is comprised of all distorting effects both systematic and random and is not limited to inaccuracy of a transfer function.

2.7 Calibration Error

Calibration error is inaccuracy permitted by a manufacturer when a sensor is calibrated in the factory. This error is of a systematic nature, meaning that it is added to all possible real transfer functions. It shifts the accuracy of transduction for each stimulus point by a constant. This error is not necessarily uniform over the range and may change depending on the type of error in calibration. For example, let us consider a two-point calibration of a real linear transfer function (thick line in Fig. 2.10). To determine the slope and the intercept of the function, two stimuli, s_1 and s_2, are applied to the sensor. The sensor responds with two corresponding

[4]The flow rate can be measured in foot per minute (fpm).

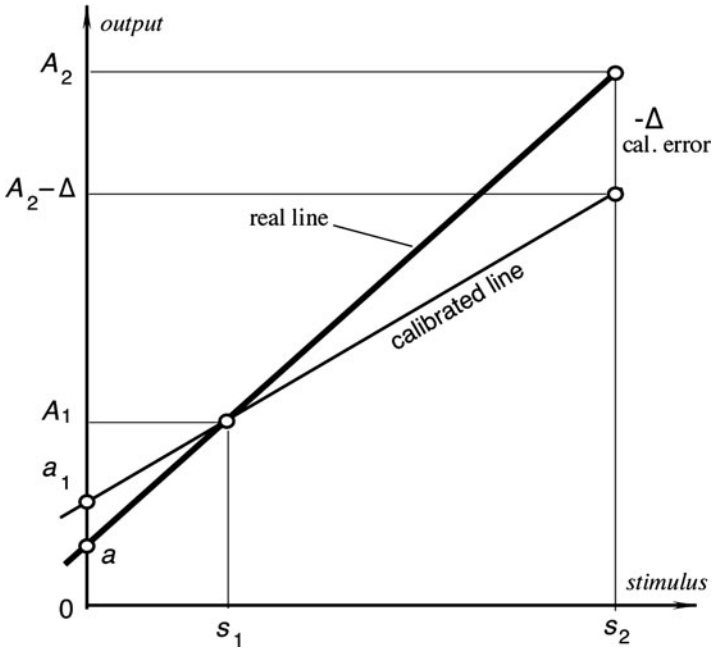

Fig. 2.10 Calibration error

output signals A_1 and A_2. Let us say the first response was measured absolutely accurately, while the other response was measured with error –Δ. This results in errors in the slope and intercept calculation. A new erroneous intercept, a_1 will differ from the true intercept, a, by

$$\delta_a = a_1 - a = \frac{\Delta}{s_2 - s_1}, \tag{2.30}$$

and the slope will be calculated with error:

$$\delta_b = -\frac{\Delta}{s_2 - s_1} \tag{2.31}$$

2.8 Hysteresis

A hysteresis error is a deviation of the sensor's output at a specified point of the input signal when it is approached from the opposite directions (Fig. 2.11). For example, a displacement sensor when the object moves from left to right at a certain point produces voltage, which differs by 20 mV from that when the object moves

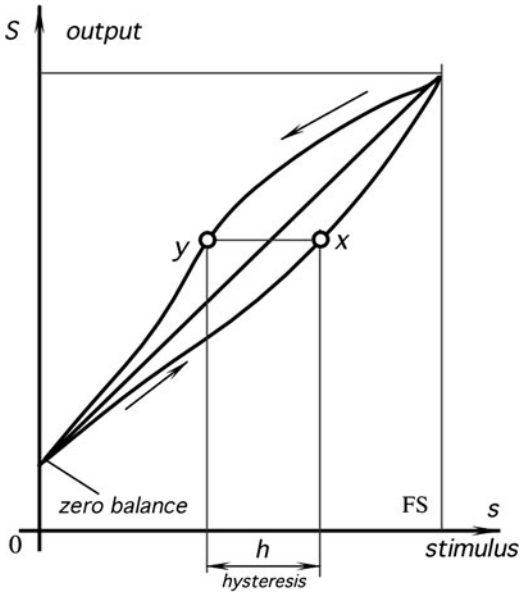

Fig. 2.11 Transfer function with hysteresis

from right to left. If sensitivity of the sensor is 10 mV/mm, the hysteresis error in terms of displacement units is 2 mm. Typical causes for hysteresis are geometry of design, friction, and structural changes in the materials.

2.9 Nonlinearity

Nonlinearity error is specified for sensors whose transfer function may be approximated by a straight line (2.2). A nonlinearity is a maximum deviation (L) of a real transfer function from the approximation straight line. The term "linearity" actually means "nonlinearity." When more than one calibration run is made, the worst linearity seen during any one calibration cycle should be stated. Usually, it is specified either in % of span or in terms of measured value, for instance, in kPa or °C. "Linearity," when not accompanied by a statement explaining what sort of straight line it is referring to, is meaningless. There are several ways to specify nonlinearity, depending how the line is superimposed on the transfer function. One way is to use *terminal* points (Fig. 2.12a), that is, to determine output values at the smallest and highest stimulus values and to draw a straight line through these two points (line 1). Here, near the terminal points, the nonlinearity error is the smallest and it is higher somewhere in between.

In some applications, higher accuracy may be desirable in a particular narrower section of the input range. For instance, a medical thermometer should have the best accuracy in a fever definition region which is between 37°C and 38°C. It may have

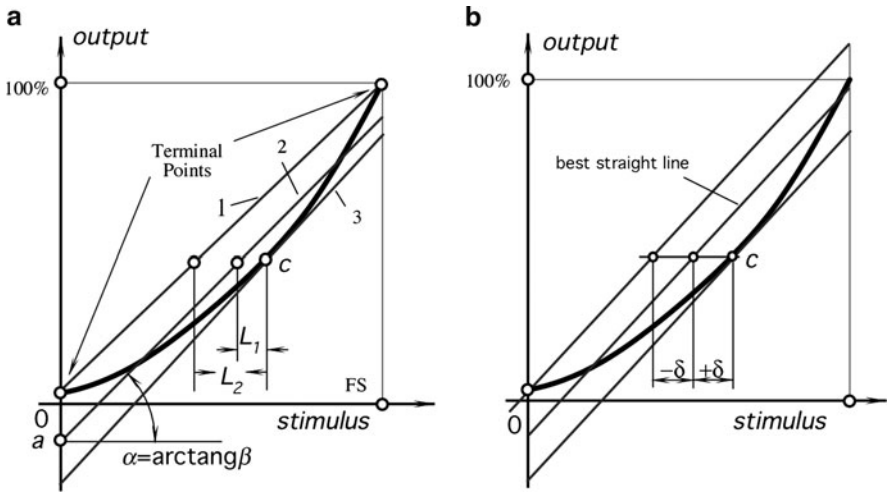

Fig. 2.12 Linear approximations of a nonlinear transfer function (**a**); and independent linearity; (**b**)

a somewhat lower accuracy beyond these limits. Usually, such a sensor is calibrated in the region where the highest accuracy is desirable. Then, the approximation line may be drawn through the calibration point *c* (line 3 in Fig. 2.12a). As a result, nonlinearity has the smallest value near the calibration point and it increases toward the ends of the span. In this method, the line is often determined as tangent to the transfer function in point *c*. If the actual transfer function is known, the slope of the line can be found from (2.5).

Independent linearity is referred to the so-called "best straight line" (Fig. 2.12b), which is a line midway between two parallel straight lines closest together and enveloping all output values on a real transfer function.

Depending on the specification method, approximation lines may have different intercepts and slopes. Therefore, nonlinearity measures may differ quite substantially from one another. A user should be aware that manufacturers often publish the smallest possible number to specify nonlinearity, without defining what method was used.

2.10 Saturation

Every sensor has its operating limits. Even if it is considered linear, at some levels of the input stimuli, its output signal no longer will be responsive. Further increase in stimulus does not produce a desirable output. It is said that the sensor exhibits a span-end nonlinearity or saturation (Fig. 2.13).

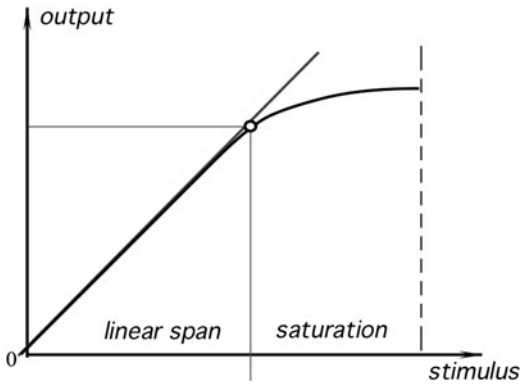

Fig. 2.13 Transfer function with saturation

2.11 Repeatability

Repeatability (reproducibility) error is caused by the inability of a sensor to represent the same value under presumably identical conditions. The repeatability is expressed as a maximum difference between the output readings as determined by two calibrating cycles (Fig. 2.14a), unless otherwise specified. It is usually represented as % of FS:

$$\delta_{\mathrm{r}} = \frac{\Delta}{\mathrm{FS}} 100\% \tag{2.32}$$

Possible sources of the repeatability error may be thermal noise, build up charge, material plasticity, etc.

2.12 Dead Band

Dead band is insensitivity of a sensor in a specific range of the input signals (Fig. 2.14b). In that range, the output may remain near a certain value (often zero) over an entire dead band zone.

2.13 Resolution

Resolution describes smallest increments of stimulus, which can be sensed. When a stimulus continuously varies over the range, the output signals of some sensors will not be perfectly smooth, even under the no-noise conditions. The output may change in small steps. This is typical for potentiometric transducers, occupancy

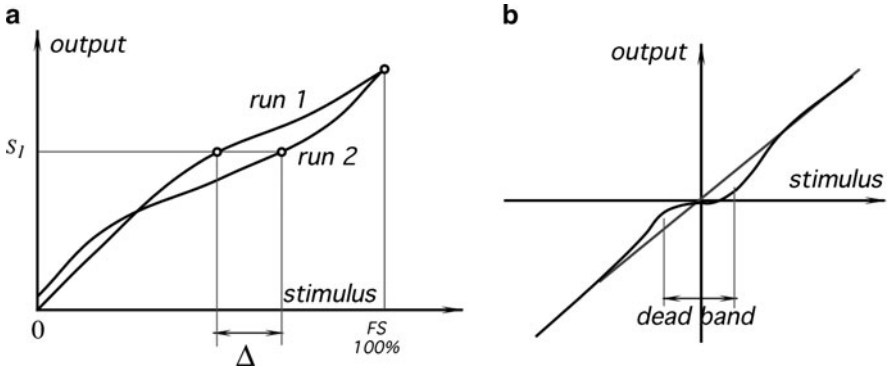

Fig. 2.14 Repeatability error (**a**). The same output signal S_1 corresponds to two different input signals Dead-band zone in a transfer function (**b**)

infrared detectors with grid masks, and other sensors where the output signal change is enabled only upon a certain degree of stimulus variation. Besides, any signal that is converted into a digital format is broken into small steps where a number is assigned to each step. The magnitude of the input variation, which results in the output's smallest step, is specified as resolution under specified conditions (if any). For instance, for the occupancy detector the resolution may be specified as follows: "resolution – minimum equidistant displacement of the object for 20 cm at a 5 m distance." For wire-wound potentiometric angular sensors, resolution may be specified as "a minimum angle of 0.5°." Sometimes, it may be specified as percents of full scale (FS). For instance, for the angular sensor having 270° FS, the 0.5° resolution may be specified as 0.181% of FS. It should be noted that the step size may vary over the range, hence, the resolution may be specified as typical, average, or "worst." The resolution of digital output format sensors is given by the number of bits in the data word. For instance, the resolution may be specified as "8-bit resolution." This statement to make sense must be accomplished with either of FS value or the value of LSB (least significant bit). When there are no measurable steps in the output signal, it is said that the sensor has continuous or infinitesimal resolution (sometimes erroneously referred to as "infinite resolution").

2.14 Special Properties

Special input properties may be needed to specify for some sensors. For instance, light detectors are sensitive within a limited optical bandwidth. Therefore, it is appropriate to specify for them a spectral response.

2.15 Output Impedance

Output impedance Z_{out} is important to know to better interface a sensor with the electronic circuit. The output impedance is connected to the input impedance Z_{in} of the circuit either in parallel (voltage connection) or in series (current connection). Figure 2.15 shows these two connections. The output and input impedances generally should be represented in a complex form, as they may include active and reactive components. To minimize the output signal distortions, a current generating sensor (B) should have an output impedance as high as possible while the circuit's input impedance should be low. For the voltage connection (A), a sensor is preferable with lower Z_{out} while the circuit should have Z_{in} as high as practical.

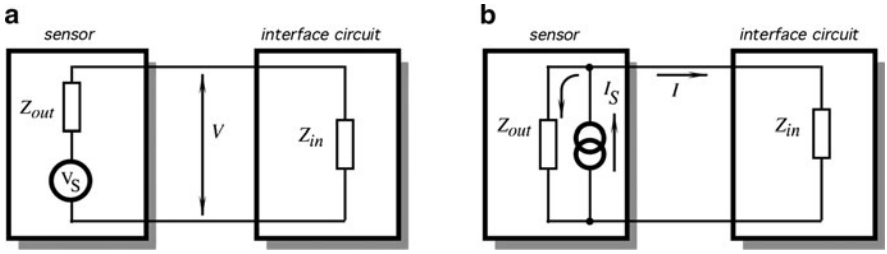

Fig. 2.15 Sensor connection to an interface circuit sensor has voltage output (**a**), sensor has current output (**b**)

2.16 Output Format

Output format is a set of the output electrical characteristics that is produced by the sensor alone or together with the excitation circuit. The characteristics may include voltage, current, charge, frequency, amplitude, phase, polarity, shape of a signal, time delay, and digital code. Figure 2.16 shows examples of the output electrical signals in form of current or voltage. A sensor manufacturer should provide sufficient information on the output format to allow for efficient applications.

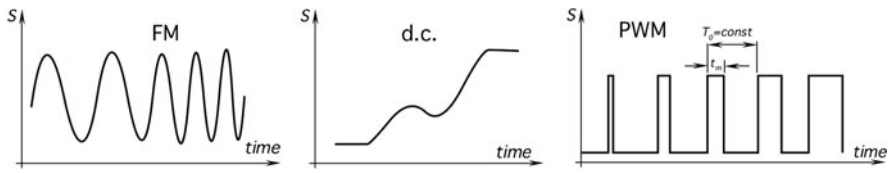

Fig. 2.16 Examples of output signals: sine wave constant amplitude with frequency modulation (FM), analog signal (d.c.) changing within the output range, and pulse-width modulation (PWM) of rectangular pulses of constant period but variable width

2.17 Excitation

Excitation is the electrical signal needed for operation of an active sensor. Excitation is specified as a range of voltage and/or current. For some sensors, the frequency and shape of the excitation signal and its stability must also be specified. Spurious variations in the excitation may alter the sensor transfer function and cause output errors.

An example of excitation signal specification is as follows:

Maximum current through a thermistor	In still air	50 μA
	In water	1 mA

2.18 Dynamic Characteristics

Under static conditions (a very slow changing input stimulus) a sensor is fully described by its transfer function, span, calibration, etc. However, when an input stimulus varies with an appreciable rate, a sensor response generally does not follow with perfect fidelity. The reason is that both the sensor and its coupling with the source of stimulus cannot always respond instantly. In other words, a sensor may be characterized with a time-dependent characteristic, which is called a dynamic characteristic. If a sensor does not respond instantly, it may represent the stimulus as somewhat different from the real, that is, the sensor responds with a dynamic error. A difference between a static and dynamic error is that the latter is always time-dependent. If a sensor is part of a control system, which has its own dynamic characteristics, the combination may cause at best a delay in representing a true value of a stimulus or at worst-cause spurious oscillations.

Warm-up time is the time between applying to the sensor power or excitation signal and the moment when the sensor can operate within its specified accuracy. Many sensors have a negligibly short warm-up time. However, some detectors, especially those that operate in a thermally controlled environment (e.g., a thermostat) may require seconds and minutes of warm-up time before they are fully operational within the specified accuracy limits.

In a control system theory, it is common to describe the input-output relationship through a constant-coefficient linear differential equation. Then, sensor's dynamic (time-dependent) characteristics can be studied by evaluating such an equation. Depending on the sensor design, the differential equation can be of several orders.

A zero-order sensor is characterized by a transfer function that is time independent. Such a sensor does not incorporate any energy storage devices, like a capacitor. A zero-order sensor responds instantaneously. In other words, such a sensor does not need any dynamic characteristics to be specified.

A first-order differential equation describes a sensor that incorporates one energy storage component. The relationship between the input $s(t)$ and output $S(t)$ is differential equation

$$b_1 \frac{\mathrm{d}S(t)}{\mathrm{d}t} + b_0 S(t) = s(t) \tag{2.33}$$

A typical example of a first-order sensor is a temperature sensor where energy storage is thermal capacity.

The first-order sensors may be specified by a manufacturer in various ways. A typical is a frequency response, which specifies how fast a first-order sensor can react to a change in the input stimulus. The frequency response is expressed in Hz or rad/s to specify the relative reduction in the output signal at certain frequency (Fig. 2.17a). A commonly used reduction number (frequency limit) is –3 dB. It shows at what frequency the output voltage (or current) drops by about 30%. Frequency response limit f_u is often called the upper cutoff frequency, as it is considered the highest frequency, which a sensor can process.

The frequency response directly relates to a speed response, which is defined in units of input stimulus per unit of time. How to specify, frequency or speed, in any particular case depends on the sensor type, its application, and the preference of a designer.

Another way to specify speed response is by time, which is required by the sensor to reach 90% of a steady-state or maximum level upon exposure to a step stimulus. For the first-order response, it is very convenient to use a so-called time constant. Time constant τ is a measure of the sensor's inertia. In electrical terms, it is equal to a product of electrical capacitance and resistance: $\tau = CR$. In thermal terms, thermal capacity and thermal resistances should be used instead. Practically, the time constant can be easily measured.

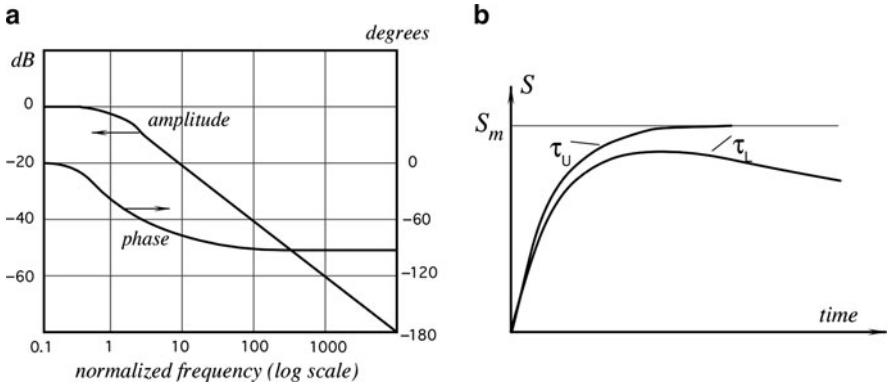

Fig. 2.17 Frequency characteristic (**a**) and response of a first-order sensor (**b**) with limited upper and lower cutoff frequencies τ_u and τ_L are the corresponding time constants

A first order system response is as follows:

$$S = S_m(1 - e^{-t/\tau}), \tag{2.34}$$

where S_m is steady-state output, t is time, and e is base of natural logarithm.

Substituting $t = \tau$, we get

$$\frac{S}{S_m} = 1 - \frac{1}{e} = 0.6321 \tag{2.35}$$

In other words, after an elapse of time equal to one time constant, the response reaches about 63% of its steady-state level. Similarly, it can be shown that after two time constants, the height will be 86.5% and after three time constants it will be 95% of the level that would be reached at infinite time.

Cutoff frequency shows what is the lowest or highest frequency of stimulus the sensor can process. The upper cutoff frequency shows how fast the sensor reacts, the lower cutoff frequency shows how slowly changing stimuli the sensor can process. Figure 2.17b depicts the sensor's response when both upper and lower cutoff frequencies are limited. As a rule of thumb, a simple formula can be used to establish a connection between the cutoff frequency f_c (either upper and lower) and time constant in a first-order sensor:

$$f_c \approx \frac{0.159}{\tau} \tag{2.36}$$

Phase shift at a specific frequency defines how the output signal lags behind in representing the stimulus change (Fig. 2.17a). The shift is measured in angular degrees or rads and is usually specified for a sensor that processes periodic signals. If a sensor is part of a feedback control system, it is very important to know its phase characteristic. The phase lag reduces the phase margin of the system and may result in overall instability.

A second-order differential equation describes a sensor that incorporates two energy storage components. The relationship between the input $s(t)$ and output $S(t)$ is differential equation

$$b_2 \frac{d^2S(t)}{dt^2} + b_1 \frac{dS(t)}{dt} + b_0 S(t) = s(t) \tag{2.37}$$

An example of a second-order sensor is an accelerometer that incorporates an inertial mass and a spring (Fig. 2.18).

A second-order response is specific for a sensor that responds with a periodic signal. Such a periodic response may be very brief and we say that the sensor is damped, or it may be of a prolonged time and even may oscillate continuously. Naturally, for a sensor such a continuous oscillation is a malfunction and must be avoided. Any second-order sensor may be characterized by a resonant (natural) frequency, which is a number expressed in Hz or rad/s. The natural frequency

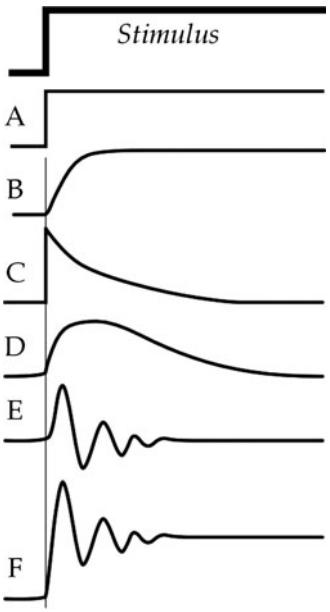

Fig. 2.18 Types of responses unlimited upper and lower frequencies (A); first-order limited upper cutoff frequency (B); first-order limited lower cutoff frequency (C); first-order limited both upper and lower cutoff frequencies (D); narrow bandwidth response (resonant) (E); wide bandwidth with resonant (F)

shows where the sensor's output signal increases considerably. When the sensor behaves as if the output conforms to the standard curve of a second-order response, the manufacturer will state the natural frequency and the damping ratio of the sensor. The resonant frequency may be related to mechanical, thermal, or electrical properties of the detector. Generally, the operating frequency range for the sensor should be selected well below (at least 60%) or above the resonant frequency. However, in some sensors, the resonant frequency is the operating point. For instance, in glass breakage detectors (used in security systems) the resonant makes the sensor selectively sensitive to a narrow bandwidth, which is specific for the acoustic spectrum, produced by shuttered glass.

Damping is the progressive reduction or suppression of the oscillation in the sensor having higher than the first-order response. When the sensor's response is as fast as possible but without an overshot, the response is said to be critically damped (Fig. 2.19). Underdamped response is when the overshot occurs and the over-damped response is slower than the critical. The damping ratio is a number expressing the quotient of the actual damping of a second-order linear transducer by its critical damping.

For an oscillating response, as shown in Fig. 2.19, a damping factor is a measure of damping, expressed (without sign) as the quotient of the greater by the littlest of pair of consecutive swings in opposite directions of the output signal, about an ultimately steady-state value. Hence, the damping factor can be measured as:

$$\text{Damping factor} = \frac{F}{A} = \frac{A}{B} = \frac{B}{C} = \text{etc} \tag{2.38}$$

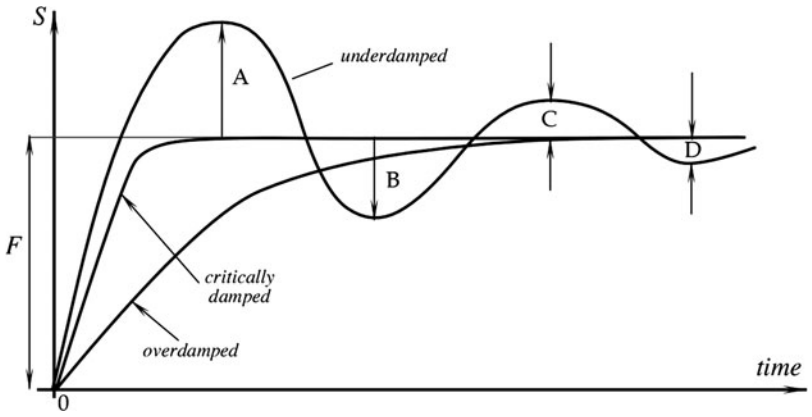

Fig. 2.19 Responses of sensors with different damping characteristics

2.19 Environmental Factors

Storage conditions are nonoperating environmental limits to which a sensor may be subjected during a specified period without permanently altering its performance under normal operating conditions. Usually, storage conditions include the highest and the lowest storage temperatures and maximum relative humidities at these temperatures. Words "noncondensing" may be added to the relative humidity number. Depending on the sensor's nature, some specific limitation for the storage may need to be considered. For instance, maximum pressure, presence of some liquids, gases or contaminating fumes.

Short- and long-term stabilities (drift) are parts of the accuracy specification. The short-term stability is manifested as changes in the sensor's performance within minutes, hours, or even days. Eventually, it is another way to express repeatability (see above) as drift may be bi-directional. That is, the sensor's output signal may increase or decrease, which is, in other terms, may be described as ultralow frequency noise. The long-term stability may be related to aging of the sensor materials, which is an irreversible change in the material's electrical, mechanical, chemical, or thermal properties. That is, the long-term drift is usually unidirectional. It happens over a relatively long time span, such as months and years. Long-term stability is one of the most important for the sensors that are used for precision measurements. Aging greatly depends on environmental storage and operating conditions, how well the sensor components are isolated from the environment and what materials are used for their fabrication. Aging phenomenon is typical for sensors having organic components and, in general, is not an issue for a sensor made with only nonorganic materials. For instance, glass-coated metal–oxide thermistors exhibit much greater long-term stability as compared with epoxy coated.

A powerful way to improve long-term stability is to preage the component at extreme conditions. The extreme conditions may be cycled from the lowest to the

highest. For instance, a sensor may be periodically swung from freezing to hot temperatures. Such accelerated aging not only enhances stability of the sensor's characteristics, but also improves the reliability (see below), as the preaging process reveals many hidden defects. For instance, thermistors may be greatly improved if they are maintained at +150°C for a month before they are calibrated and installed into a product.

Environmental conditions to which a sensor is subjected do not include variables that the sensor measures. For instance, an air pressure sensor usually is subjected not just to air pressure, but to other influences as well, such as temperatures of air and surrounding components, humidity, vibration, ionizing radiation, electromagnetic fields, gravitational forces, etc. All these factors may and usually do affect the sensor's performance. Both static and dynamic variations in these conditions should be considered. Some environmental conditions usually are of a multiplicative nature, that is, they alter a transfer function of the sensor, for instance changing its gain. One example is resistive strain gauge whose sensitivity increases with temperature.

Environmental stability is quite broad and usually a very important requirement. Both the sensor designer and the application engineer should consider all possible external factors, which may affect the sensor's performance. A piezoelectric accelerometer may generate spurious signals if affected by a sudden change in ambient temperature, electrostatic discharge, formation of electrical charges (triboelectric effect), vibration of a connecting cable, electromagnetic interferences (EMI), etc. Even if a manufacturer does not specify such effects, an application engineer should simulate them during the prototype phase of the design process. If, indeed, the environmental factors degrade the sensor's performance, additional corrective measures may be required (see Chap. 5). For instance, placing the sensor in a protective box, electrical shielding, using a thermal insulation, or a thermostat.

Temperature factors are very important for sensor performance, they must be known, and accounted for. The operating temperature range is the span of ambient temperatures given by their upper and lower extremes (e.g., "–20 to +100°C") within which the sensor maintains its specified accuracy. Many sensors change with temperature and their transfer functions may shift significantly. Special compensating elements are often incorporated either directly into the sensor or into signal conditioning circuits to compensate for temperature errors. The simplest way of specifying tolerances of thermal effects is provided by the error-band concept, which is simply the error band that is applicable over the operating temperature band. A temperature band may be divided into sections while the error band is separately specified for each section. For example, a sensor may be specified to have an accuracy of ±1% in the range from 0°C to 50°C, ±2% from –20°C to 0°C, and from +50°C to 100°C, and ±3% beyond these ranges within operating limits which are specified from –40°C to +150°C.

Temperatures will also affect dynamic characteristics, particularly when they employ viscous damping. A relatively fast temperature change may cause the sensor to generate a spurious output signal. For instance, a dual pyroelectric sensor in a motion detector is insensitive to slow varying ambient temperature. However,

when the temperature changes fast, the sensor will generates electric current, which may be recognized by a processing circuit as a valid response to a stimulus, thus causing a false positive detection.

A self-heating error may be specified when an excitation signal is absorbed by a sensor and changes its temperature by such a degree that it may affect its accuracy. For instance, a thermistor temperature sensor requires passage of electric current, causing heat dissipation within the sensor's body. Depending on its coupling with the environment, the sensors' temperature may increase due to a self-heating effect. This will result in errors in temperature measurement since the thermistor now acts as an additional spurious source of thermal energy. The coupling depends on the media where the sensor operates – a dry contact, liquid, air, etc. The worst coupling may be through still air. For thermistors, manufacturers often specify self-heating errors in air, stirred liquid, or other media.

A sensor's temperature increase above its surroundings may be found from formula:

$$\Delta t^{\circ} = \frac{V}{R(\xi vc + \alpha)} \tag{2.39}$$

where ζ is the sensor's mass density, c is specific heat, v is the volume of the sensor, α is the coefficient of thermal coupling between the sensor and the outside (thermal conductivity), R is the electrical resistance of the sensor, and V is the effective constant voltage across the resistance. If a self-heating results in an error, (2.39) may be used as a design guidance. For instance, to increase α, a thermistor detector should be well coupled to the object by increasing the contact area, applying thermally conductive grease or using thermally conductive adhesives. Also, high resistance sensors and low measurement voltages are preferable.

2.20 Reliability

Reliability is the ability of a sensor to perform a required function under stated conditions for a stated period. It is expressed in statistical terms as a probability that the device will function without failure over a specified time or a number of uses. It should be noted that reliability is not a characteristic of drift or noise stability. It specifies a failure, that is, temporary or permanent, exceeding the limits of a sensor's performance under normal operating conditions.

Reliability is an important requirement, however, it is rarely specified by the sensor manufacturers. Probably, the reason for that is the absence of a commonly accepted measure for the term. In the United States, for many electronic devices, the procedure for predicting in-service reliability is the MTBF (mean-time-between-failure) calculation described in MIL-HDBK-217 standard. Its basic approach is to arrive at a MTBF rate for a device by calculating the individual

failure rates of the individual components used and by factoring in the kind of operation the device will see: its temperature, stress, environmental, and screening level (measure of quality). Unfortunately, MTBF reflects reliability only indirectly and it is often hardly applicable to everyday use of the device. The qualification tests on sensors are performed at combinations of the worst possible conditions. One approach (suggested by MIL-STD-883) is 1,000 h, loaded at maximum temperature. This test does not qualify for such important impacts as fast temperature changes. The most appropriate method of testing would be accelerated life qualification. It is a procedure that emulates the sensor's operation, providing real-world stresses, but compressing years into weeks. Three goals are behind the test: to establish MTBF; to identify first failure points that can then be strengthened by design changes; and to identify the overall system practical life time.

One possible way to compress time is to use the same profile as actual operating cycle, including maximum loading and power-on, power-off cycles, but expanded environmental highest and lowest ranges (temperature, humidity, and pressure). The highest and lowest limits should be substantially broader than normal operating conditions. Performance characteristics may be outside specifications, but must return to those when the device is brought back to the specified operating range. For example, if a sensor is specified to operate up to 50°C at the highest relative humidity (RH) of 85% at maximum supply voltage of +15 V, it may be cycled up to 100°C at 99% RH and at +18 V power supply. To estimate number of test cycles (n), the following empirical formula (developed by Sandstrand Aerospace, Rockford, IL and Interpoint Corp., Redmond, WA) [3] may be useful:

$$n = N\left(\frac{\Delta T_{\max}}{\Delta T_{test}}\right)^{2.5} \tag{2.40}$$

where N is the estimated number of cycles per lifetime, $\Delta T_{\max}$ is the maximum specified temperature fluctuation, and ΔT_{test} maximum cycled temperature fluctuation during the test. For instance, if the normal temperature is 25°C, the maximum specified temperature is 50°C, cycling was up to 100°C, and over the life time (say, 10 years) the sensor was estimated will be subjected to 20,000 cycles, then the number of test cycles is calculated as:

$$n = 20,000 \cdot \left(\frac{50-25}{100-25}\right)^{2.5} = 1283.$$

As a result, the accelerated life test requires about 1300 cycles instead of 20,000. It should be noted, however, that the 2.5 factor was derived from a solder fatigue multiple since that element is heavily influenced by cycling. Some sensors have no solder connections at all and some might have even more sensitive to cycling substances than solder, for instance, electrically conductive epoxy. Then, the factor should be selected somewhat smaller. As a result of the accelerated life test, the reliability may be expressed as a probability of failure. For instance, if 2 out of 100

sensors (with an estimated life time of 10 years) failed the accelerated life test, the reliability is specified as 98% over 10 years. For a better understanding of accelerated life tests and accelerated aging, refer to the excellent text [4].

A sensor, depending on its application, may be subjected to some other environmental effects, which potentially can alter its performance or uncover hidden defects. Among such additional tests are as follows:

- High temperature/high humidity while being fully electrically powered. For instance, a sensor may be subjected to its maximum allowable temperature at 85–90% relative humidity (RH) and kept under these conditions during 500 h. This test is very useful for detecting contaminations and evaluation of packaging integrity. Life of a sensor that operates at normal room temperatures is often accelerated at 85°C and 85%RH. This accelerated life test sometimes is called an "85–85 test."
- Mechanical shocks and vibrations may be used to simulate adverse environmental conditions, especially in evaluation wire bonds, adhesion of epoxy, etc. A sensor may be dropped to generate high level accelerations (up to 3000*g* of force). The drops should be made on different axes. Harmonic vibrations should be applied to the sensor over the range, which includes its natural frequency. In the United States, military standard #750, methods 2016 and 2056 are often used for mechanical tests.
- Extreme storage conditions may be simulated, for instance at +100°C and –40°C while maintaining a sensor for at least 1,000 h under these conditions. This test simulates storage and shipping conditions and usually is performed on nonoperating devices. The upper and lower temperature limits must be consistent with the sensor's physical nature. For example, a TGS pyroelectric sensors manufactured in the past by Philips are characterized by a Curie temperature of +60°C. Approaching and surpassing this temperature results in a permanent destruction of the sensor sensitivity. Hence, temperature of such sensors should never exceed +50°C, which must be clearly specified and marked on its packaging material.
- Thermal shock or temperature cycling (TC) is subjecting a sensor to alternate extreme conditions. For example, it may be dwelled for 30 min at –40°C, then rapidly moved to +100°C for 30 min, and then back to cold. The method must specify total number of cycling, like 100 or 1,000. This test helps to uncover die bond, wire bond, epoxy connections, and packaging integrity.
- To simulate sea conditions, sensors may be subjected to a salt spray atmosphere for a specified time, for example 24 h. This helps to uncover its resistance to corrosion and structural defects.

2.21 Application Characteristics

Design, weight, and overall dimensions are geared to specific areas of applications.

Price may be a secondary issue when the sensor's reliability and accuracy are of paramount importance. If a sensor is intended for life support equipment, weapons

or spacecraft, a high price tag may be well justified to assure high accuracy and reliability. On the other hand, for a very broad range of consumer applications, the price of a sensor often becomes a corner stone of a design.

2.22 Uncertainty

Nothing is perfect in this world, at least in a sense that we perceive it. All materials are not exactly as we think they are. Our knowledge even of the purest of the materials is always approximate, machines are not perfect and never produce perfectly identical parts according to drawings. All components experience drifts related to the environment and their aging, external interferences may enter the system and alter its performance and modify the output signal. Workers are not consistent and the human factor is nearly always present. Manufacturers fight an everlasting battle for the uniformity and consistency of the processes, yet the reality is that every part produced is never ideal and carries an uncertainty of its properties. Any measurement system consists of many components, including sensors. Thus, no matter how accurate the measurement is, its only an approximation or estimate of the true value of the specific quantity subject to measurement, that is the stimulus or measurand. The result of measurement should be considered complete only when accompanied by a quantitative statement of its uncertainty. We simply never can be 100% sure of the measured value.

When taking individual measurements (samples) under noisy conditions we expect that stimulus s is represented by the sensor as having a somewhat different value s', so that the error in measurement is expressed as

$$\delta = s' - s, \tag{2.41}$$

The difference between the error that is specified by (2.27) and uncertainty should always be clearly understood. An error can be compensated to a certain degree by correcting its systematic component. The result of such a correction can unknowably be very close to the unknown true value of the stimulus and thus it will have a very small error. Yet, in spite of a small error, the uncertainty of measurement may be very large so we cannot really trust that the error is indeed that small. In other words, an error is what we unknowably get when we measure, while uncertainty is what we think how large that error might be.

The International Committee for Weight and Measures (CIPM) considers that uncertainty consists of many factors that can be grouped into two classes or types [1, 2]:

A: those, which are evaluated by statistical methods;
B: those, which are evaluated by other means.

This division is not clear-cut and the borderline between A and B is somewhat illusive. Generally, A components of uncertainty arise from random effects, while the B components arise from systematic effects.

Type *A* uncertainty is generally specified by a standard deviation s_i, equal to the positive square root of the statistically estimated variance s_i^2, and the associated number of degrees of freedom υ_i. For such a component the *standard* uncertainty is $u_i = s_i$. Standard uncertainty represents each component of uncertainty that contributes to the uncertainty of the measurement result.

Evaluation of a Type *A* standard uncertainty may be based on any valid statistical method for treating data. Examples are calculating standard deviation of the mean of a series of independent observations using the method of least squares to fit a curve to data in order to estimate the parameters of the curve and their standard deviations. If the measurement situation is especially complicated, one should consider obtaining the guidance of a statistician.

Evaluation of a Type *B* of standard uncertainty is usually based on scientific judgment using all the relevant information available, such may include

- Previous measurement data
- Experience with or general knowledge of the behavior and property of relevant sensors, materials, and instruments
- Manufacturer's specifications
- Data obtained during calibration and other reports and
- Uncertainties assigned to reference data taken from handbooks and manuals

For detailed guidance of assessing and specifying standard uncertainties one should consult specialized texts, for instance [5].

When both *A* and *B* uncertainties are evaluated, they should be combined to represent the combined standard uncertainty. This can be done by using a conventional method for combining standard deviations. This method is often called the law of propagation of uncertainty and in common parlance is known as "root-sum-of-squares" (square root of the sum-of-the-squares) or "RSS" method of combining uncertainty components estimated as standard deviations:

$$u_c = \sqrt{u_1^2 + u_2^2 + \cdots + u_i^2 + \cdots + u_n^2}, \tag{2.42}$$

where n is a number of standard uncertainties in the uncertainty budget.

Table 2.4 shows an example of the uncertainty budget for an electronic thermometer with a thermistor sensor, which measures temperature of a water bath. While compiling such a table one shall be very careful not to oversee any standard uncertainty not only in a sensor, but also in the interface instrument, experimental setup, and the object of measurement. This shall be done for various environmental conditions, which may include temperature, humidity, atmospheric pressure, power supply variations, transmitted noise, aging, and many other factors.

No matter how accurately any individual measurement is made, that is, how close the measured temperature is to the true temperature of an object, one never can be sure that it is indeed accurate. The combined standard uncertainty of 0.068°C does not mean that error of measurement is no greater than 0.068°C. That value is just a standard deviation and if an observer has enough patience he may find that

Table 2.4 Uncertainty budget for thermistor thermometer

Source of uncertainty	Standard uncertainty (°C)	Type
Calibration of sensor	0.03	*B*
Measured Errors		
Repeated observations	0.02	*A*
Sensor noise	0.01	*A*
Amplifier noise	0.005	*A*
Sensor aging	0.025	*B*
Thermal loss through connecting wires	0.015	*A*
Dynamic error due to sensor's inertia	0.005	*B*
Temperature instability of object of measurement	0.04	*A*
Transmitted noise	0.01	*A*
Misfit of transfer function	0.02	*B*
Ambient drifts		
Voltage reference	0.01	*A*
Bridge resistors	0.01	*A*
Dielectric absorption in A/D capacitor	0.005	*B*
Digital resolution	0.01	*A*
Combined standard uncertainty	0.068	

individual errors may be much larger. The word "uncertainty" by its very nature implies that the uncertainty of the result of a measurement is an estimate and generally does not have well-defined limits.

References

1. Kelley CT (2003) Solving nonlinear equations with Newton's method, No. 1 Fundamentals of Algorithms. SIAM, Philadelphia, PA
2. Süli E, Mayers D (2003) An introduction to numerical analysis. Cambridge University Press, Cambridge
3. Stoer J, Bulirsch R (2002) Introduction to numerical analysis. Springer, Berlin, pp 93–106
4. Suhir E (2007) How to make a device into a product. Accelerated life testing (ALT), its role attributes, challenges, pitfalls and interaction with qualification tests. In: Suhir E, Lee YC, Wong CP (eds) Micro- and opto-electronic materials and structures: physics, mechanics, design, reliability, packaging, vol 2, chap 8. Springer, Berlin, pp 203–230
5. Better reliability via system tests. Electronic engineering times, CMP Publication, pp 40–41, Aug. 19, 1991
6. CIPM (1981) BIPM Proc.-Verb. Com. Int. Poids et Mesures 49, pp 8–9, No. 26 (in French)
7. ISO guide to the expression of uncertainty in measurements (1993) International Organization for Standardization, Geneva, Switzerland
8. Taylor BN, Kuyatt CE (1994) Guidelines for evaluation and expressing the uncertainty of NIST measurement results. NIST Technical Note 1297. US Government Printing Office, Washington DC

Chapter 3
Physical Principles of Sensing

The way we have to describe Nature
is generally incomprehensible to us.
– Richard P. Feynman,
QED. The Strange Theory of Light and Matter
It should be possible to explain
the laws of physics to a barmaid.
– Albert Einstein

Since a sensor is a converter of generally nonelectrical effects into electrical signals, one and often several transformation steps are required before the electric output signal can be generated. These steps involve changes of types of energy where the final step must produce electrical signal of a desirable format. As it was mentioned in Chap. 1, generally there are two types of sensors: direct and complex. A direct sensor is the one that can directly convert a nonelectrical stimulus into electric signal. Many stimuli cannot be directly converted into electricity, thus multiple conversion steps would be required. If, for instance, one wants to detect displacement of an opaque object, a fiber optic sensor can be employed. A pilot (excitation) light is generated by a light emitting diode (LED), transmitted via an optical fiber to the object and reflected from its surface. The reflected photon flux enters the receiving optical fiber and propagates toward a photodiode where it produces an electric current representing the distance from the fiber optic end to the object. We see that such a sensor involves transformation of electrical current into photons, propagation of photons through some refractive media, reflection, and conversion back into electric current. Therefore, such a sensing process includes two energy conversion steps and a manipulation of the optical signal as well.

There are several physical effects resulting from a direct generation of electrical signals in response to nonelectrical influences and thus can be used in direct sensors. Examples are thermoelectric (Seebeck) effect, piezoelectricity, and photo-effect.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_3, © Springer Science+Business Media, LLC 2010

This chapter examines various physical effects which can be used for a direct conversion of stimuli into electric signals. Since all such effects are based on fundamental principles of physics, we briefly review these principles from the stand point of sensor technologies.

3.1 Electric Charges, Fields, and Potentials

There is a well-known phenomenon to those who live in dry climates – the possibility of the generation of sparks by friction involved in walking across the carpet. This is a result of the so-called triboelectric effect,[1] which is a process of an electric charge separation due to object movements, friction of clothing fibers, air turbulence, atmosphere electricity, etc. There are two kinds of charges. Like charges repel and the unlike charges attract each other. Benjamin Franklin (1706–1790), among his other remarkable achievements, was the first American physicist. He named one charge *negative* and the other *positive*. These names have remained to this day. He made an elegant experiment with a kite flying in a thunderstorm to prove that the atmospheric electricity is of the same kind as produced by friction. In doing the experiment, Franklin was extremely lucky, as several Europeans who were trying to repeat his test were severely injured by the lightning and one was killed.[2]

A triboelectric effect is a result of a mechanical charge redistribution. For instance, rubbing a glass rod with silk strips off electrons from the surface of the rod, thus leaving an abundance of positive charges, i.e., giving the rod a positive charge. It should be noted that the electric charge is conserved; it is neither created nor destroyed. Electric charges can be only moved from one place to another. Giving negative charge means taking electrons from one object and placing them onto another (charging it negatively). The object that loses some amount of electrons is said gets a positive charge.

A triboelectric effect influences an extremely small number of electrons as compared with the total electronic charge in an object. The actual amount of charges in any object is very large. To illustrate this, let us consider total number if electrons in a U.S. copper penny[3] [1]. The coin weighs 3.1 g, therefore, it can be shown that the total number of atoms in it is about 2.9×10^{22}. A copper atom has a positive nuclear charge of 4.6×10^{-18} C and, respectively, the same electronic charge of the opposite polarity. A combined charge of all electrons in a penny is $q = (4.6 \times 10^{-18}\ \text{C/atom})\ (2.9 \times 10^{22}\ \text{atoms}) = 1.3 \times 10^{5}$ C, a very large charge indeed.

[1] The prefix *tribo-* means "pertinent to friction".

[2] A Russian physicist of German extraction Georg Wilhelm Richmann (1711–1753) was killed in St. Petersburg during a thunderstorm experiment when a ball lightning having a size of a fist jumped from the electrometer and struck him in a forehead.

[3] Now, the U.S. pennies are just copper-plated (2.5%), but till 1982 they contained 95% of copper.

This electronic charge from a single copper penny may generate sufficient current of 0.91 A to operate a 100 W light bulb for 40 h.

With respect to electric charges, there are three kinds of materials: conductors, isolators, and semiconductors. In conductors, electric charges (electrons) are free to move through the material, whereas in isolators they are not. Although there is no perfect isolator, the isolating ability of fused quartz is about 10^{25} times as great as that of copper, so that for practical purposes many materials are considered perfect isolators. The semiconductors are intermediate between conductors and isolators in their ability to conduct electricity. Among the elements, silicon and germanium are well-known examples. In semiconductors, the electrical conductivity may be greatly increased by adding small amounts of other elements: traces of arsenic or boron are often added to silicon for this purpose.

Figure 3.1a shows an object which carries a positive electric charge q. If a small positive electric test charge q_0 is positioned in the vicinity of a charged object, it will be subjected to a repelling electric force. If we place a negative charge on the object, it will attract the test charge. In a vector form, the repelling (or attracting) force is shown as **f**. The bold face indicates a vector notation. A fact that the test charge is subjected to force without a physical contact between charges means that the volume of space occupied by the test charge may be characterized by a so-called electric field.

The electric field in each point is defined through the force as

$$\mathbf{E} = \frac{\mathbf{f}}{q_0} \tag{3.1}$$

Here **E** is vector of the same direction as **f** because q_0 is scalar. Formula (3.1) expresses an electric field as a force divided by a property of a test charge. The test charge must be very small not to disturb the electric field. Ideally, it should be infinitely small, however, since the charge is quantized, we cannot contemplate a free test charge whose magnitude is smaller than the electronic charge: $e = 1.602 \times 10^{-19}$ C.

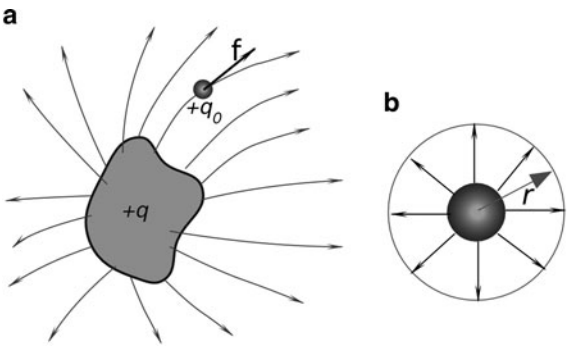

Fig. 3.1 Positive test charge in vicinity of a charged object (**a**) and electric field of a spherical object (**b**)

The field is indicated in Fig. 3.1a by the field lines, which in every point of space are tangent to the vector of force. By definition, the field lines start on the positive plate and end on the negative. The density of field lines indicates the magnitude of electric field **E** in any particular volume of space.

For a physicist, any field is a physical quantity, which can be specified simultaneously for all points within a given region of interest. Examples are pressure field, temperature fields, electric fields, and magnetic fields. A field variable may be a scalar (for instance, temperature field) or a vector (for instance, a gravitational field around the earth). The field variable may or may not change with time. A vector field may be characterized by a distribution of vectors which form the so-called flux (symbol Φ). Flux is a convenient description of many fields, such as electric, magnetic, thermal, etc. The word flux is derived from the Latin word *fluere* (to flow). A familiar analogy of flux is a stationary, uniform field of fluid flow (water) characterized by a constant flow vector **v**, the constant velocity of the fluid at any given point. In case of electric field, nothing flows in a formal sense. If we replace **v** by **E** (vector representing electric field) the field lines form flux. If we imagine a hypothetical closed surface (Gaussian surface) ***S***, a connection between the charge q and flux can be established as

$$\varepsilon_0 \Phi_E = q, \tag{3.2}$$

where $\varepsilon_0 = 8.8542 \times 10^{-12}$ C^2/Nm^2 is the permitivity constant, or by integrating flux over the surface

$$\varepsilon_0 \oint \mathbf{E} ds = q, \tag{3.3}$$

where the integral is equal to Φ_E. In the above equations, known as Gauss' law, charge q is the net charge surrounded by the Gaussian surface. If a surface encloses equal and opposite charges, the net flux Φ_E is zero. The charge outside the surface makes no contribution to the value of q, nor does the exact location of the inside charges affect this value. Gauss' law can be used to make an important prediction: *An exact charge on an insulated conductor is in equilibrium, entirely on its outer surface*. This hypothesis was shown to be true even before either Gauss' law or Coulomb law was advanced. The Coulomb law itself can be derived from the Gauss' law. It states that the force acting on a test charge is inversely proportional to a squared distance from the charge

$$f = \frac{1}{4\pi\varepsilon_0} \frac{qq_0}{r^2}. \tag{3.4}$$

Another result of Gauss' law is that the electric field outside any spherically symmetrical distribution of charge (Fig. 3.1b) is directed radially and has magnitude (note that magnitude is not a vector)

$$E = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}, \tag{3.5}$$

where r is the distance from the sphere's center.

Similarly, the electric field inside a uniform sphere of charge q is directed radially and has magnitude

$$E = \frac{1}{4\pi\varepsilon_0}\frac{qr}{R^3}, \tag{3.6}$$

where R is the sphere's radius and r is the distance from the sphere's center. It should be noted that the electric field in the center of the sphere ($r = 0$) is equal to zero.

If the electric charge is distributed along an infinite (or, for the practical purposes, long) line (Fig. 3.2a), the electric field is directed perpendicularly to the line and has the magnitude

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}, \tag{3.7}$$

where r is the distance from the line and λ is the linear charge density (charge per unit length).

The electric field due to an infinite sheet of charge (Fig. 3.2b) is perpendicular to the plane of the sheet and has magnitude

$$E = \frac{\sigma}{2\varepsilon_0}, \tag{3.8}$$

where σ is the surface charge density (charge per unit area). However, for an isolated conductive object, the electric field is two times stronger

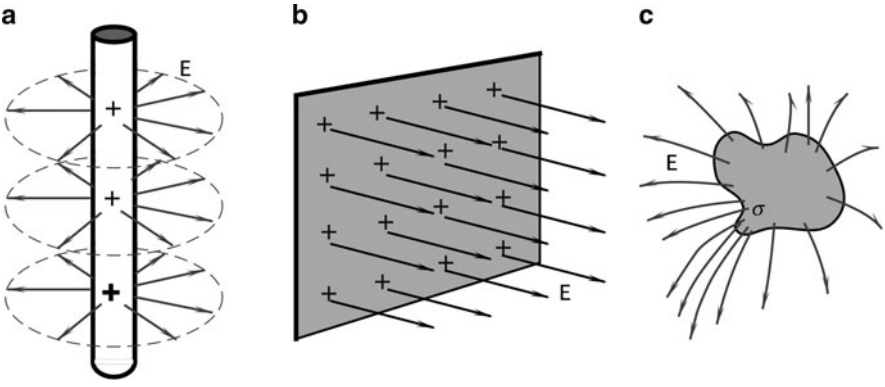

Fig. 3.2 Electric field around an infinite line (**a**) and near an infinite sheet (**b**). A pointed conductor concentrates an electric field (**c**)

$$E = \frac{\sigma}{\varepsilon_0}. \tag{3.9}$$

The apparent difference between electric fields is a result of different geometries – the former is an infinite sheet and the latter is an object of an arbitrary shape. A very important consequence of Gauss' law is that electric charges are distributed only on the outside surface. This is a result of repelling forces between charges of the same sign; all charges try to move as far as possible from one another. The only way to do this is to move to the foremost distant place in the material, which is the outer surface. Of all places on the outer surface the most preferable places are the areas with the highest curvatures. This is why pointed conductors are the best concentrators of the electric field Fig. 3.2c). A very useful scientific and engineering tool is a Faraday cage, a room entirely covered by either grounded conductive sheets or a metal net. No matter how strong the external electric field, it will be essentially zero inside the cage. This makes metal airplanes, cars, and ships the best protectors during thunderstorms, because they act as virtual Faraday cages. It should be remembered however that the Faraday cage, while being a perfect shield against electric fields, is of little use to protect against magnetic fields, unless it is made of a thick ferromagnetic material.

An electric dipole is a combination of two opposite charges which are placed at a distance $2a$ apart Fig. 3.3a). Each charge will act on a test charge with force which defines electric fields $\mathbf{E_1}$ and $\mathbf{E_2}$ produced by individual charges. A combined electric field of a dipole, $\mathbf{E}$, is a vector sum of two fields. The magnitude of the field is

$$E = \frac{1}{4\pi\varepsilon_0}\frac{qa}{r^3}, \tag{3.10}$$

where r is the distance from the center of the dipole. The essential properties of the charge distribution are magnitude of the charge q and the separation $2a$. In formula (3.10) charge and distance are entered only as a product. This means that, if we

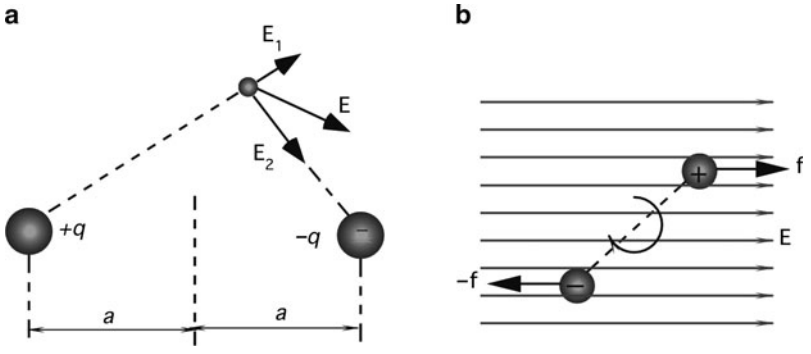

Fig. 3.3 Electric dipole (**a**); an electric dipole in an electric field is subjected to a rotating force (**b**)

measure E at various distances from the electric dipole (assuming that distance is much longer than a), we can never deduce q and $2a$ separately, but only the product $2qa$. For instance, if q is doubled and a is cut into half, the electric field will not change. The product $2qa$ is called the electric dipole moment p. Thus, (3.10) can be rewritten as

$$E = \frac{1}{4\varepsilon_0}\frac{p}{r^3}. \tag{3.11}$$

The spatial position of a dipole may be specified by its moment in a vector form: $\boldsymbol{p}$. Not all materials have a dipole moment: gases such as methane, acetylene, ethylene, carbon dioxide, and many others have no dipole moment. On the other hand, carbon monoxide has a weak dipole moment (0.37×10^{-30} Cm) and water has a strong dipole moment (6.17×10^{-30} Cm).

Dipoles are found in crystalline materials and form a foundation for such sensors as piezo and pyroelectric detectors. When a dipole is placed in an electric field, it becomes subjected to a rotation force (Fig. 3.3b). Usually, a dipole is part of a crystal which defines its initial orientation. An electric field, if strong enough, will align the dipole along its lines. Torque which acts on a dipole in a vector form is

$$\boldsymbol{\tau} = \boldsymbol{pE}. \tag{3.12}$$

Work must be done by an external agent to change the orientation of an electric dipole in an external electric field. This work is stored as potential energy U in the system consisting of the dipole and the arrangement used to set up the external field. In a vector form this potential energy is

$$\boldsymbol{U} = -\boldsymbol{pE}. \tag{3.13}$$

A process of dipole orientation is called poling. The aligning electric filed must be strong enough to overcome a retaining force in the crystalline stricture of the material. To ease this process, the material during the poling is heated to increase mobility of its molecular structure. The poling of a ceramic or crystalline polymer is used in fabrication of piezo- and pyroelectric materials.

The electric field around the charged object can be described not only by the vector $\boldsymbol{E}$, but by a scalar quantity, the electric potential V as well. Both quantities are intimately related and usually it is a matter of convenience which one to use in practice. A potential is rarely used as a description of an electric field in a specific point of space. A potential difference (voltage) between two points is the most common quantity in electrical engineering practice. To find the voltage between two arbitrary points, we may use the same technique as above, a small positive test charge q_0. If the electric charge is positioned in point A, it stays in equilibrium being under influence of force $q_0\boldsymbol{E}$. It may remain there theoretically infinitely long. Now, if we try to move it to another point B, we have to work against the electric field.

Work $-W_{AB}$, which is done against the field (that is why it has a negative sign) to move the charge from A to B defines voltage between these two points

$$V_B - V_A = -\frac{W_{AB}}{q_0}. \tag{3.14}$$

Correspondingly, the electrical potential at point B is smaller than at point A. The SI unit for voltage is 1V = 1 joule/coulomb. For convenience, point A is chosen to be very far away from all charges (theoretically at an infinite distance) and the electric potential at that point is considered to be zero. This allows us to define electric potential at any other point as

$$V = -\frac{W}{q_0}. \tag{3.15}$$

This equation tells us that the potential near the positive charge is positive, because moving the positive test charge from infinity to the point in a field, must be made against a repelling force. This will cancel the negative sign in formula (3.15). It should be noted that the potential difference between two points is independent of a path at which the test charge is moving. It is strictly a description of the electric field difference between the two points. If we travel through the electric field along a straight line and measure V as we go, the rate of change of V with distance l that we observe is the components of $\boldsymbol{E}$ in that direction

$$E_l = -\frac{dV}{dl}. \tag{3.16}$$

The minus sign tells us that $\boldsymbol{E}$ points in the direction of decreasing V. As it follows from (3.16), the appropriate unit for electric field is volts/meter (V/m).

3.2 Capacitance

Let us take two isolated conductive objects of arbitrary shape (plates) and connect them to the opposite poles of a battery (Fig. 3.4a). The plates will receive equal amounts of opposite charges. That is, a negatively charged plate will receive additional electrons while there will be a deficiency of electrons in the positively charged plate. Now, let us disconnect the battery. If the plates are totally isolated and exist in a vacuum, they will remain charged theoretically infinitely long. A combination of plates which can hold an electric charge is called a capacitor. If a small positive electric test charge, q_0, is positioned between the charged objects, it will be subjected to an electric force from the positive plate to the negative. The positive plate will repel the positive test charge and the negative will attract it, thus resulting in a combined push-pull force. Depending on the position of the test charge between the oppositely charged objects, the force will have a specific magnitude and direction which is characterized by vector **f**.

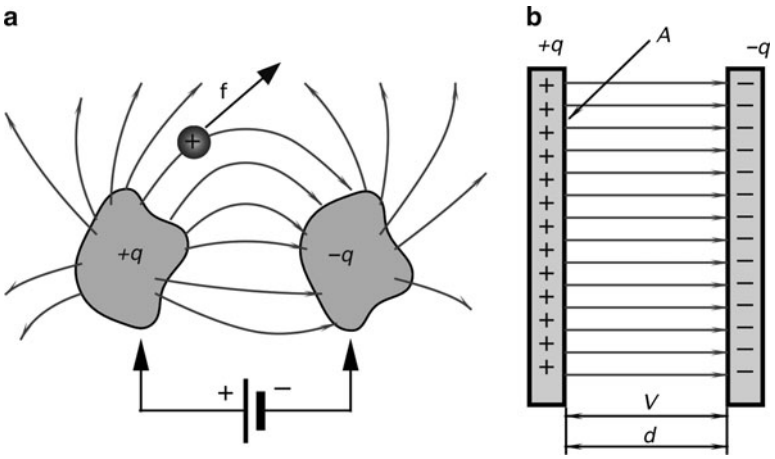

Fig. 3.4 Electric charge and voltage define capacitance between two objects (**a**); parallel-plate capacitor (**b**)

The capacitor may be characterized by q, the magnitude of charge on either conductor, shown in Fig. 3.4a, and by V, the positive potential difference between the conductors. It should be noted that q is not a net charge on the capacitor, which is zero. Further, V is not the potential of either plate, but the potential difference between them. The ratio of charge to voltage is constant for each capacitor:

$$\frac{q}{V} = C. \tag{3.17}$$

This fixed ratio, C, is called the capacitance of the capacitor. Its value depends on the shapes and relative position of the plates. The ratio C also depends on the medium in which the plates are immersed. Note, that C is always positive since we use the same sign for both q and V. The SI unit for capacitance is 1 farad = 1 coulomb/volt, which is represented by the abbreviation F. A farad is a very large capacitance, hence, in practice submultiples of the farad are generally used:

1 picofarad (pF)	$= 10^{-12}$ F
1 nanofarad (nF)	$= 10^{-9}$ F
1 microfarad (μF)	$= 10^{-6}$ F

When connected into an electronic circuit, capacitance may be represented as a "complex resistance":

$$\frac{V}{i} = -\frac{1}{j\omega C}. \tag{3.18}$$

where $j = \sqrt{-1}$ and i is the sinusoidal current having a frequency of ω, meaning that the complex resistance of a capacitor drops at higher frequencies. This is called Ohm's law for the capacitor. The minus sign and complex argument indicate that the voltage across the capacitor lags by 90° behind the current.

Capacitance is a very useful physical phenomenon in a sensor designer's toolbox. It can be successfully applied to measure distance, area, volume, pressure, force, chemical composition, etc. The following background establishes fundamental properties of the capacitor and gives some useful equations. Figure 3.4b shows a parallel-plate capacitor in which the conductors take the form of two plane parallel plates of area A separated by a distance d. If d is much smaller that the plate dimensions, the electric field between the plates will be uniform, which means that the field lines (lines of force $\boldsymbol{f}$) will be parallel and evenly spaced. The laws of electromagnetism requires that there be some "fringing" of the lines at the edges of the plates, but for small enough d, we can neglect it for our present purpose.

3.2.1 Capacitor

To calculate the capacitance we must relate V, the potential difference between the plates, to q, the capacitor charge (3.17)

$$C = \frac{q}{V}. \tag{3.19}$$

Alternatively, the capacitance of a flat capacitor in vacuum can be found from

$$C = \frac{\varepsilon_0 A}{d}. \tag{3.20}$$

In a capacitive sensor, the value of capacitance is the measure of a stimulus, so to change the capacitance, the stimulus needs to change one of the parameters that define the capacitance. These parameters are established by the key formula (3.20). It establishes a relationship between the plate area and distance between the plates. Varying one of them will change the capacitor's value, which can be measured quite accurately by an appropriate circuit. It should be noted that the above equations hold only for capacitors of the parallel type. A change in geometry will require modified formulas. A ratio, A/d, may be called a geometry factor for a parallel-plate capacitor.

A cylindrical capacitor, which is shown in Fig. 3.5a, consists of two coaxial cylinders of radii a and b, and length l. For the case when $l \gg b$, we can ignore fringing effects and calculate capacitance from the following formula

$$C = \frac{2\pi\varepsilon_0 l}{\ln\frac{b}{a}}. \tag{3.21}$$

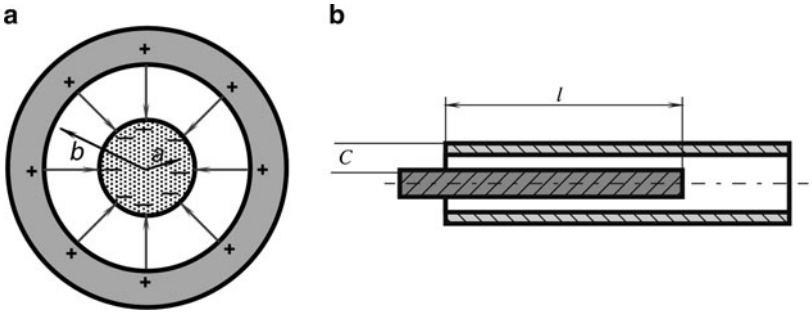

Fig. 3.5 Cylindrical capacitor (**A**); capacitive displacement sensor (**B**)

In this formula l has a meaning of length of the overlapping conductors (Fig. 3.5b) and $2\pi l(\ln b/a)^{-1}$ may be called a geometry factor for a coaxial capacitor. A useful displacement sensor can be built with such a capacitor if the inner conductor can be moved in and out of the outer conductor. According to (3.21), the capacitance of such a sensor is in a linear relationship with the displacement, l.

3.2.2 Dielectric Constant

Equation (3.20) holds for a parallel-plate capacitor with its plates in vacuum (or air, for most practical purposes). In 1837, Michael Faraday first investigated the effect of completely filling the space between the plates with a dielectric. He had found that the effect of the filling is to increase the capacitance of the device by a factor of k, which is known as the dielectric constant of the material.

The increase in capacitance due to the dielectric presence is a result of molecular polarization. In some dielectrics (for instance, in water), molecules have a permanent dipole moment, while in other dielectrics, molecules become polarized only when an external electric field is applied. Such polarization is called induced. In both cases, either permanent electric dipoles or acquired by induction, tend to align molecules with an external electric field. This process is called dielectric polarization. It is illustrated by Fig. 3.6a, which shows permanent dipoles before and Fig. 3.6b after an external electric field is applied to the capacitor. In the former case, there is no voltage between the capacitor plates and all dipoles are randomly oriented. After the capacitor is charged, the dipoles will align with the electric field lines, however, thermal agitation will prevent a complete alignment. Each dipole forms its own electric field, which is predominantly is oppositely directed with the external electric field, $\mathbf{E}_0$. Due to a combined effect of a large number of dipoles ($\mathbf{E}'$), the electric filed in the capacitor becomes weaker ($\mathbf{E} = \mathbf{E_0} + \mathbf{E}'$) when the field, $\mathbf{E_0}$, would be in the capacitor without the dielectric.

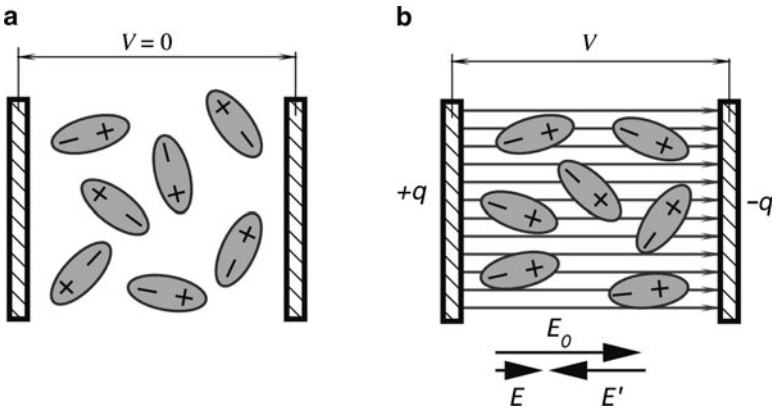

Fig. 3.6 Polarization of dielectric

Reduced electric field leads to a smaller voltage across the capacitor: $V = V_o/\kappa$. Substituting it into formula (3.19) we get an expression for the capacitor with dielectric

$$C = \kappa \frac{q}{V_0} = \kappa C_0. \tag{3.22}$$

For the parallel plate capacitor we thus have

$$C = \frac{\kappa \varepsilon_0 A}{d}. \tag{3.23}$$

In a more general form, the capacitance between two objects may be expressed through a geometry factor, G

$$C = \varepsilon_0 \kappa G. \tag{3.24}$$

were G depends on the shape of the objects (plates) and their separation. Thus, formula (3.24) establishes that a capacitance can also be modulated by varying the dielectric constant κ. Table A.5 (Appendix) gives dielectric constants, κ, for different materials.

Dielectric constants must be specified for test frequency and temperature. Some dielectrics have a very uniform dielectric constant over a broad frequency range (for instance, polyethylene), while others display strong negative frequency dependence, that is, a dielectric constant decreases with frequency. Temperature dependence is also negative. Figure 3.7 illustrates κ for water as a function of temperature.

In a "good" capacitor used in electronic circuits, a dielectric constant κ and geometry G better be stable. Ideally, they should not vary with temperature, humidity, pressure, or any other environmental factors. "Good" capacitors are critical factors that determine quality of the electronic circuits. However, if you

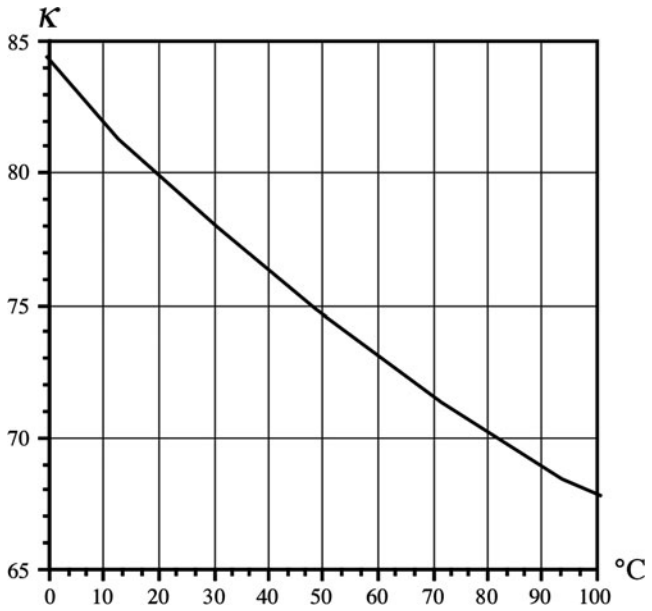

Fig. 3.7 Dielectric constant of water as a function of temperature

want to design a capacitive sensor, you need to make a "bad" capacitor, whose value varies with temperature, or humidity, or pressure, or whatever you need to sense. By allowing a capacitor's parameter to vary *selectively* with a specific stimulus, one can build a useful sensor.

Let's consider a capacitive water level sensor (Fig. 3.8a). The sensor is fabricated in form of a coaxial capacitor where the surface of each conductor is coated with a thin isolating layer to prevent an electric short circuit through water (the isolator is a dielectric which we disregard in the following analysis because it does not change in the process of measurement). The sensor is immersed in a water tank. When the level increases, water fills more and more space between the sensor's coaxial conductors, thus changing the average dielectric constant between the conductors and, according to (3.24), subsequently changing the sensor's capacitance. Total capacitance of the coaxial sensor is

$$C_h = C_1 + C_2 = \varepsilon_0 G_1 + \varepsilon_0 \kappa G_2, \tag{3.25}$$

where C_1 is the capacitance of the water-free portion of the sensor and C_2 is the capacitance of the water-filled portion. The corresponding geometry factors are designated G_1 and G_2.

From formulas (3.21) and (3.25), total sensor capacitance can be found as

$$C_h = \frac{2\pi\varepsilon_0}{\ln\frac{b}{a}}[H - h(1-\kappa)], \tag{3.26}$$

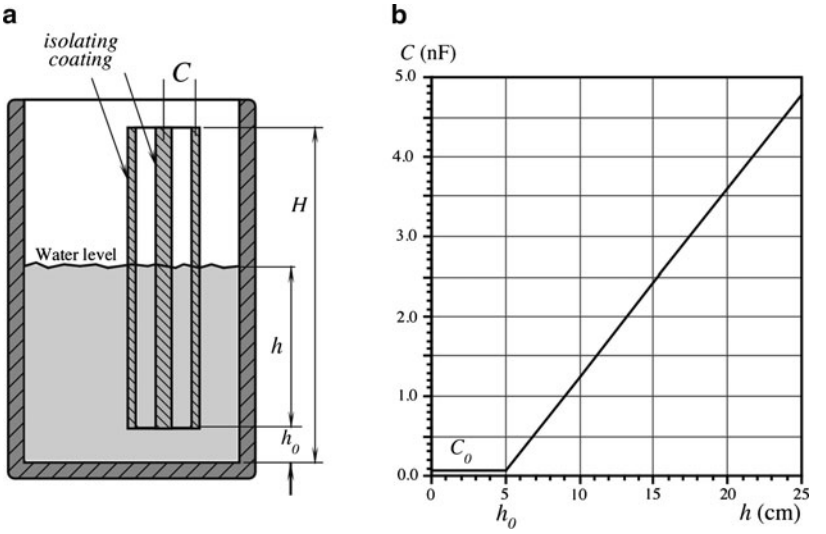

Fig. 3.8 Capacitive water level sensor (**A**); capacitance as function of the water level (**B**) (sensor's dimensions are as follows: a=10 mm, b = 12 mm, H = 200 mm, liquid—water)

where h is height of the water-filled portion of the sensor. If the water is at or below the level h_0, the capacitance remains constant because $h = 0$.

$$C_0 = \frac{2\pi\varepsilon_o}{\ln\frac{b}{a}}H. \tag{3.27}$$

Figure 3.8b shows the water level-capacitance dependence. It is a straight line from level h_0. Since the dielectric constant of water is temperature-dependent (Fig. 3.7) the capacitive water level sensor shall be combined with a temperature sensor, for instance, a thermistor or RTD, which would monitors water temperature for compensating the temperature dependence. The appropriate temperature correction may be performed by the electronic signal conditioner.

The slope of the transfer function (3.26) line depends on the liquid. For instance, if instead of water the sensor measures level of transformer oil, it is expected to be 22 times less sensitive (see Table A.5).

Another example of a capacitive sensor is a humidity sensor. In such a sensor, a dielectric between the capacitor plates is fabricated of a material that is hygroscopic, that is, it can absorb water molecules. Material dielectric constant varies with the amount of absorbed moisture. According to (3.24), this changes the capacitance that can be measured and converted to a value of relative humidity. Figure 3.9 illustrates the dependence between a capacitance and a relative humidity of such a sensor. The dependence is not perfectly linear but this usually can be taken care of during the signal processing.

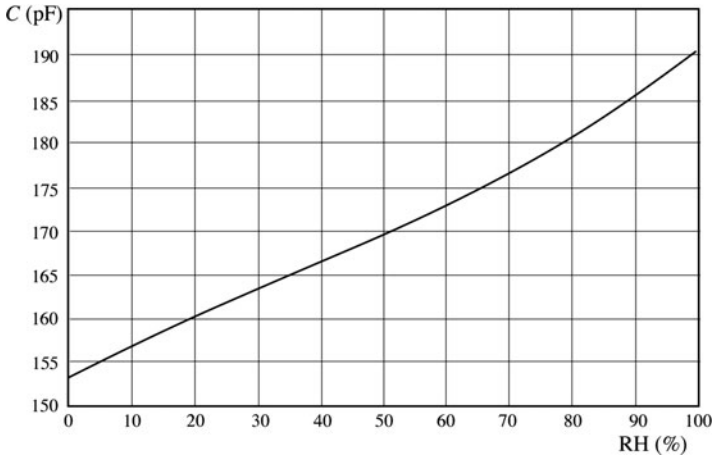

Fig. 3.9 Transfer function of a capacitive relative humidity sensor

3.3 Magnetism

Magnetic properties were discovered in prehistoric times in certain specimens of an iron ore mineral known as magnetite (Fe_3O_4). It was also discovered that pieces of soft iron that rubbed against a magnetic material acquired the same property of acting as a magnet, i.e., attracting other magnets and pieces of iron. The first comprehensive study of magnetism was made by William Gilbert. His greatest contribution was his conclusion that the earth acts as a huge magnet. The word magnetism comes from the district of Magnesia in Asia Minor, which is one of the places at which the magnetic stones were found.

There is a strong similarity between electricity and magnetism. One manifestation of this is that two electrically charged rods have like and unlike ends, very much in the same way as two magnets have opposite ends. In magnets, these ends are called S (south) and N (north) poles. The like poles repeal and the unlike attract. Contrary to electric charges, the magnetic poles always come in pairs. This is proven by breaking magnets into any number of parts. Each part, no matter how small, will have a north pole and a south pole. This suggests that the cause of magnetism is associated with atoms or their arrangements or, more probably, with both.

If we place a magnetic pole in a certain space, that space about the pole appears to have been altered from what it was before. To demonstrate this, bring into that space a piece of iron. Now, it will experience a force that it will not experience if the magnet is removed. This altered space is called a magnetic field. The field is considered to exert a force on any magnetic body brought into the field. If that magnetic body is a small bar magnet or a magnetic needle, the magnetic field will be found to have direction. By definition, the direction of this field at any point is given by the direction of the force exerted on a small unit north pole. Directions of field lines are by definition from north to south pole. Figure 3.10a shows the

direction of the field by arrows. A tiny test magnet is attracted in the direction of the force vector **F**. Naturally, about the same force but of opposite direction is exerted on the south pole of the test magnet.

The above description of the magnetic field was made for a permanent magnet. However, the magnetic field does not change its nature if it is produced by a different device: electric current passing through a conductor. It was Hans Christian Oersted (Ørsted), a Danish professor of physics, who in 1820 discovered that a magnetic field could exist where there were no magnets at all. In a series of experiments in which he was using an unusually large voltaic pile (battery) so as to produce a large current, he happened to note that a compass in the near vicinity was behaving oddly. Further investigation showed that the compass needle always oriented itself at right angles to the current carrying wire, and that it reversed its direction if either current was reversed, or the compass was changed from a position below the wire to the one above (Fig. 3.10b). Stationary electric charges make no effect on a magnetic compass (in this experiment, a compass needle is used as a tiny test magnet). It was clear that the moving electric charges were the cause of the magnetic field. It can be shown that magnetic field lines around a wire are circular and their direction depends on the direction of electric current, that is, moving electrons (Fig. 3.11). Above and below the wire, magnetic field lines are pointed in the opposite direction. That's why the compass needle turns around when it is placed below the wire.

A fundamental property of magnetism is that moving electric charges (electric current) essentially produce a magnetic field. Knowing this, Albert Einstein came up with explanation of the nature of a permanent magnet. A simplified model of a magnetic field origination process is shown in Fig. 3.12a. An electron continuously spins in an eddy motion around the atom. The electron movement constitutes a circular electric current around the atomic nucleus. That current is a cause for a small magnetic field. In other words, a spinning electron forms a permanent magnet

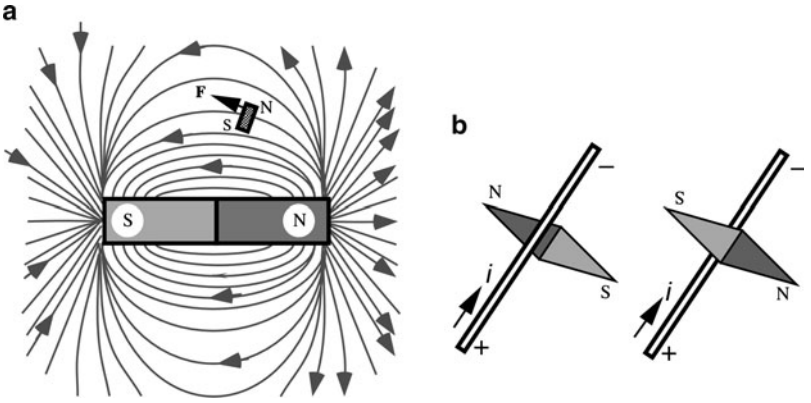

Fig. 3.10 Test magnet in a magnetic field (**a**); compass needle rotates in accordance with the direction of the electric current (**b**)

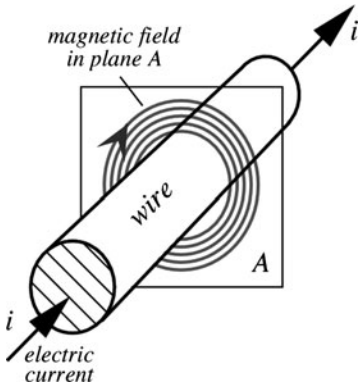

Fig. 3.11 Electric current sets a circular magnetic field around a conductor

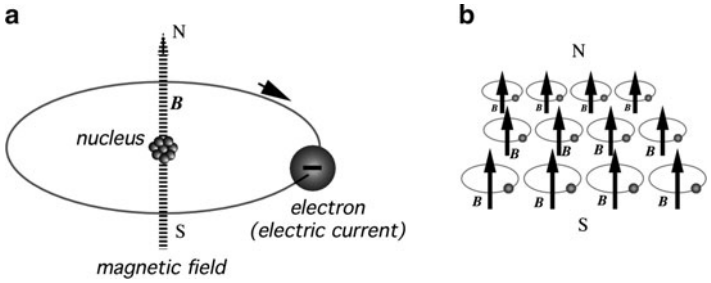

Fig. 3.12 Moving electron sets a magnetic field (**a**); superposition of field vectors results in a combined magnetic field of a magnet (**b**)

of atomic dimensions. Now, let us imagine that many of such atomic magnets are aligned in an organized fashion (Fig. 3.12b), so that their magnetic fields add up. The process of magnetization then becomes quite obvious – nothing is added or removed from the material - only orientation of atoms is made. The atomic magnets may be kept in the aligned position in some materials, which have an appropriate chemical composition and a crystalline structure. Such materials are called ferromagnetics.

3.3.1 Faraday Law

Michael Faraday pondered the question, "If an electric current is capable of producing magnetism, is it possible that magnetism can be used to produce electricity?" It took him nine or ten years to discover how. If an electric charge is moved across a magnetic field, a deflecting force is acting on that charge. It must be emphasized that it is not important what actually moves either the charge or the

source of the magnetic field. What matters is a relative displacement of those. A discovery that a moving electric charge can be deflected as a result of its interaction with the magnetic field is a fundamental in electromagnetic theory. Deflected electric charges result in an electric field generation, which, in turn, leads to a voltage difference in a conducting material, thus producing an electric current.

The intensity of a magnetic field at any particular point is defined by vector **B**, which is tangent to a magnetic field line at that point. For the better visual representation, the number of field lines per unit cross-sectional area (perpendicular to the lines) is proportional to the magnitude of **B**. Where the lines are close together, **B** is large and where they are far apart, **B** is small.

The flux of magnetic field can be defined as

$$\Phi_B = \oint B \mathrm{d}s, \tag{3.28}$$

where the integral is taken over the surface for which $\mathbf{F_B}$ is defined.

To define the magnetic field vector **B** we use a laboratory procedure where a positive electric charge q_0 is used as a test object. The charge is projected through the magnetic field with velocity **V**. A sideways deflecting force $\mathbf{F_B}$ acts on the charge (Fig. 3.13a). By "sideways" we mean that $\mathbf{F_B}$ is at a right angle to **V**. It is interesting to note that vector **V** changes its direction while moving through the magnetic field. This results in a spiral rather than parabolic motion of the charge (Fig. 3.13b). The spiral movement is a cause for a magnetoresistive effect, which forms a foundation for the magnetoresistive sensors. Deflecting force $\mathbf{F_B}$ is proportional to charge, velocity, and magnetic field

$$\mathbf{F_B} = q_0 \mathbf{V} \mathbf{B}, \tag{3.29}$$

Vector $\mathbf{F_B}$ is always at right angles to the plane formed by **v** and **B** and thus is always at right angles to **v** and to ***B***, that is why it is called a sideways force. The magnitude of magnetic deflecting force according to the rules for vector products, is

$$F_B = q_0 v B \sin \phi, \tag{3.30}$$

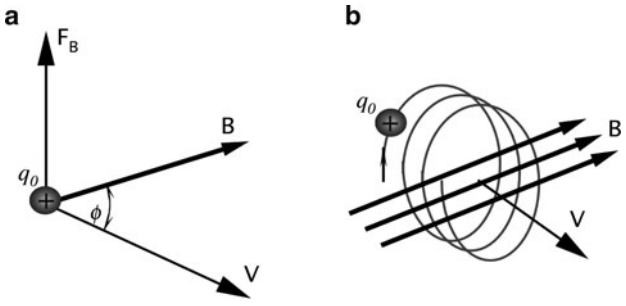

Fig. 3.13 Positive charge projected through a magnetic field is subjected to a sideways force (A); spiral movement of an electric charge in a magnetic field (B)

where Φ is the angle between vectors **V** and **B**. The magnetic force vanishes if **V** is parallel to **B**. Equation (3.30) is used for the definition of the magnetic field in terms of deflected charge, its velocity, and deflecting force. Therefore, the units of B is (Newton/coulomb)/(meter/second). In the system SI it is given name tesla (abbreviated T). Since coulomb/second is an ampere, we have 1T = 1 newton/(ampere·meter) or 1 weber per square meter. An older unit for B still is sometimes in use. It is the gauss: 1 tesla = 10^4 gauss.

3.3.2 Solenoid

A practical device to produce a magnetic field is called a solenoid. It is a long wire wound in a close-packed helix and carrying a current i. In the following discussion we assume that the helix is very long as compared with its diameter. The solenoid magnetic field is the *vector sum* of the fields setup by all the turns that make up the solenoid.

If a coil (solenoid) has widely spaced turns, the fields tends to cancel between the wires. At points inside the solenoid and reasonably far from the wires, **B** is parallel to the solenoid axis. In the limiting case of adjacent very tightly packed wires (Fig. 3.14a), the solenoid becomes essentially a cylindrical current sheet. If we apply Ampere's law to that current sheet, the magnitude of magnetic field inside the solenoid becomes

$$B = \mu_0 i_0 n, \tag{3.31}$$

where n is the number of turns per unit length and i_0 is the current through the solenoid wire. Although, this formula was derived for an infinitely long solenoid, it holds quite well for actual solenoids for internal points near the center of the solenoid. It should be noted that B does not depend on the diameter or the length of the solenoid and that B is constant over the solenoid cross-section. Since the solenoid's diameter is not a part of the equation, multiple layers of winding can be used to produce a magnetic field of higher strength. It should be noted that magnetic field outside of a solenoid is weaker than that of the inside.

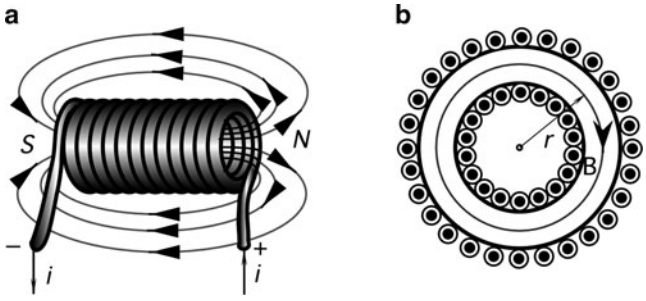

Fig. 3.14 Solenoid (**a**) and toroid (**b**)

3.3.3 Toroid

Another useful device that can produce a magnetic field is a toroid (Fig. 3.14b), which we can describe as a solenoid bent into the shape of a doughnut. A calculation of the magnetic field inside the toroid gives the following relationship

$$B = \frac{\mu_0}{2\pi} \frac{i_0 N}{r}, \tag{3.32}$$

where N is the total number of turns and r is the radius of the inner circular line where magnetic field is calculated. In contrast to a solenoid, B is not constant over the cross-section of a toroid. Besides, for an ideal case, the magnetic field is equal to zero outside a toroid.

The density of a magnetic field, or the number of magnetic lines passing through a given surface, is defined as the magnetic flux $\mathbf{\Phi}_\mathrm{B}$ for that surface

$$\Phi_B = \int \mathbf{B}\, ds \tag{3.33}$$

The integral is taken over the surface and if the magnetic field is constant and is everywhere at a right angle to the surface, the solution of the integral is very simple: $\mathbf{\Phi}_\mathrm{B} = \mathbf{B}A$, where A is the surface area. Flux, or flow of the magnetic field, is analogous to flux of electric field. The SI unit for magnetic flux, as follows from the above, is tesla·meter^2, which is named *weber*. It is abbreviated as Wb:

$$1\mathrm{Wb} = 1\mathrm{Tm}^2 \tag{3.34}$$

3.3.4 Permanent Magnets

Permanent magnets are useful components to fabricate magnetic sensors for the detection of motion, displacement, position, etc. To select the magnet for any particular application, the following characteristics should be considered:

- Residual inductance (B) in gauss – how strong the magnet is?
- Coercive force (H) in oersteds – how well will the magnet resist external demagnetization forces?
- Maximum energy product, MEP, ($B \times H$) is gauss-oersteds times 10^6. A strong magnet that is also very resistant to demagnetization forces has a high MEP. Magnets with higher MEP are better, stronger, and more expensive.
- Temperature coefficient in %/°C shows how much B changes with temperature?

Magnets are produced from special alloys (see Appendix: Table A.6). Examples are rare earth (e.g., samarium)-cobalt alloys. These are the best magnets, however,

they are too hard for machining, and must be ground if shaping is required. Their maximum MEP is about 16×10^6. Another popular alloy is Alnico, which contains aluminum, nickel, cobalt, iron, and some additives. These magnets can be cast, or sintered by pressing metal powders in a die and heating them. Sintered Alnico is well suited to mass production. Ceramic magnets contain barium or strontium ferrite (or another element from that group) in a matrix of a ceramic material that is compacted and sintered. They are poor conductors of heat and electricity, are chemically inert, and have high value of H. Another alloy for the magnet fabrication is Cunife, which contains copper, nickel, and iron. It can be stamped, swaged, drawn, or rolled into final shape. Its MEP is about 1.4×10^6. Iron-chromium magnets are soft enough to undergo machining before the final aging treatment hardens them. Their maximum MEP is 5.25×10^6. Plastic and rubber magnets consist of barium or strontium ferrite in a plastic matrix material. They are very inexpensive and can be fabricated in many shapes. Their maximum MEP is about 1.2×10^6.

A neodymium magnet (also known as NdFeB, NIB, or Neo magnet), a type of rare-earth magnet, is a permanent magnet made from an alloy of neodymium, iron, and boron to form the $Nd_2Fe_{14}B$ tetragonal crystalline structure. This material is currently the strongest type of permanent magnet. In practice, the magnetic properties of neodymium magnets depend on the alloy composition, microstructure, and manufacturing technique employed. Neodymium magnets have very much higher coercivity and energy product, but lower Curie temperature than other types of magnets.

In the 1990s it was discovered that certain molecules containing paramagnetic metal ions are capable of storing a magnetic moment at very low temperatures. In fact, these magnets are large molecules with strong magnetic properties [2]. These magnets are called single molecule magnets (SMM). Most SMMs contain manganese, but can also be found with vanadium, iron, nickel. and cobalt clusters. Advantages of SMMs include strong residual inductance, solubility in organic solvents, and sub-nanoscale dimensions. More recently it has been found that some chain systems can also display a magnetization that persists for long times at relatively higher temperatures. These systems have been called single-chain magnets (SCM).

For selecting a permanent magnet for a practical application, one can use a helpful magnet calculator on the internet (www.kjmagnetics.com/calculator.asp).

3.4 Induction

In 1831, Michael Faraday in England and Joseph Henry in the United States discovered one of the most fundamental effects of electromagnetism: an ability of a varying magnetic field to induce electric current in a wire. It is not important how the field is produced, either by a permanent magnet or by a solenoid, the effect is the same. Electric current is generated as long as the magnetic field changes. A stationary field produces no current. Faraday's law of induction says that the induced voltage, or electromotive force (e.m.f.), is equal to the rate at which the

magnetic flux through the circuit changes. If the rate of change is in Wb/s, the e.m.f. (e) will be in volts

$$e = -\frac{d\Phi_B}{dt}. \tag{3.35}$$

The minus sign is an indication of the direction of the induced e.m.f. If varying magnetic flux is applied to a solenoid, e.m.f. appears in every turn and all these e.m.f.s must be added. If a solenoid, or other coil, is wound in such a manner as each turn has the same cross-sectional area, the flux through each turn will be the same, then the induced voltage is

$$V = -N\frac{d\Phi_B}{dt} \tag{3.36}$$

where N is the number of turns. This equation may be rewritten in a form which is of interest to a sensor designer or an application engineer

$$V = -N\frac{d(BA)}{dt} \tag{3.37}$$

The equation means that the voltage in a pick-up circuit can be produced by either changing amplitude of magnetic field (B) or area of the circuit (A). Thus, induced voltage depends on

- moving the source of the magnetic field (magnet, coil, wire, etc.);
- varying the current in the coil or wire which produces the magnetic field;
- changing the orientation of the magnetic source with respect to the pick-up circuit; and
- changing the geometry of a pick-up circuit, for instance, by stretching it or squeezing, or changing the number of turns in a coil.

If an electric current passes through a coil which is situated in close proximity with another coil, according to Faraday's law, e.m.f. in a second coil will appear. However, the magnetic field penetrates not only the second coil, but the first coil as well. Thus, the magnetic field sets e.m.f. in the same coil where it is originated. This is called self-induction and the resulting voltage is called a self-induced e.m.f. Faraday's law for a central portion of a solenoid is

$$v = -\frac{d(n\Phi_B)}{dt} \tag{3.38}$$

The number in parenthesis is called the flux linkage and is an important characteristic of the device. For a simple coil with no magnetic material in the vicinity, this value is proportional to current through coil

$$n\Phi_B = Li, \tag{3.39}$$

where L is a proportionality constant, which is called the inductance of the coil. Then, (3.38) can be rewritten as

$$v = -\frac{d(n\Phi_B)}{dt} = -L\frac{di}{dt} \tag{3.40}$$

From this equation we can define inductance as

$$L = -\frac{v}{di/dt} \tag{3.41}$$

If no magnetic material is introduced in the vicinity of an *inductor* (a device possessing inductance), the value defined by (3.41) depends only on the geometry of the device. The SI unit for inductance is the volt·second/ampere, which was named after American physicist Joseph Henry (1797–1878): 1 henry = 1 volt·second/ampere. Abbreviation for henry is H.

Several conclusions can be drawn from (3.41):

- Induced voltage is proportional to the rate of change in current through the inductor.
- Voltage is essentially zero for dc.
- Voltage increases linearly with the current rate of change.
- Voltage polarity is different for increased and decreased currents flowing in the same direction.
- Induced voltage is always in the direction which opposes the change in current.

Like capacitance, inductance can be calculated from geometrical factors. For a closely packed coil it is

$$L = \frac{n\Phi_B}{i}. \tag{3.42}$$

If n is the number of turns per unit length, the number of flux linkages in the length, l, is

$$N\Phi_B = (nl) \cdot (BA), \tag{3.43}$$

where A is the cross-sectional area of the coil. For the solenoid, $B = \mu_0 ni$, then the inductance is

$$L = \frac{N\Phi_B}{i} = \mu_0 n^2 lA \tag{3.44}$$

It should be noted that lA is the volume of a solenoid that often is called a geometry factor. Thus, having the same number of turns and changing the coil geometry, its inductance may be modulated (altered) that makes an inductive transducer possible.

When connected into an electronic circuit, inductance may be represented as a "complex resistance":

$$\frac{V}{i} = \mathrm{j}\omega L, \tag{3.45}$$

where $\mathrm{j} = \sqrt{-1}$ and i is a sinusoidal current having a frequency of $\omega = 2\pi f$, meaning that the complex resistance of an inductor increases at higher frequencies. This is called Ohm's law for an inductor. Complex notation indicates that current lags behind voltage by 90°.

If two coils are brought in the vicinity of one another, and one coil conducts electric current, the magnetic field produced by that coil interacts with electrons in the second coil and induces in it e.m.f., v_2:

$$v_2 = -M_{21}\frac{di_1}{dt}, \tag{3.46}$$

where M_{21} is the coefficient of mutual inductance between two coils. The calculation of mutual inductance is not a simple exercise and in many practical cases can be easier performed experimentally. Nevertheless, for some relatively simple combinations mutual inductance have been calculated. For a coil having N turns, which is placed around a long solenoid (Fig. 3.15a), with n turns per unit length, the mutual inductance is

$$M = \mu_0 \pi R^2 nN \tag{3.47}$$

For a coil placed around a toroid (Fig. 3.15b), the mutual inductance is defined through numbers of turns, N_1 and N_2

$$M = \frac{\mu_0 N_1 N_2 h}{2\pi} \ln\left(\frac{b}{a}\right) \tag{3.48}$$

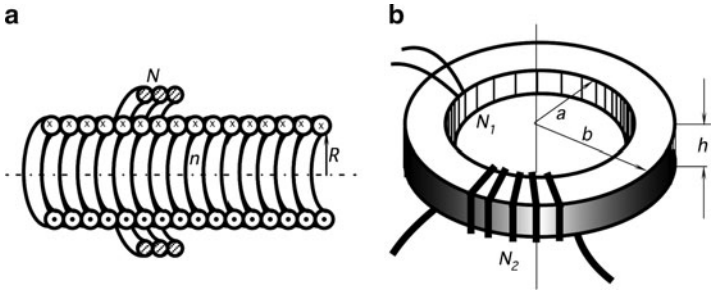

Fig. 3.15 Mutual inductances in solenoids (**a**) and in a toroid (**b**)

By varying the mutual inductance, a useful sensor can be designed. For example, a displacement sensor may have two coils, where one coil is stationary and the other is moving, causing the induced voltage to change as shown by (3.46).

3.5 Resistance

In any material, electrons move randomly like gas in a closed container. There is no preferred direction and an average concentration of electrons in any part of material is uniform (assuming that the material is homogeneous). Let us take a bar of an arbitrary material. The length of the bar is l. When the ends of the bar are connected to the battery having voltage V (Fig. 3.16), an electric field $\boldsymbol{E}$ will be setup within the material. It is easy to determine strength of the electric field

$$E = \frac{V}{l} \tag{3.49}$$

For instance, if the bar has a length of $l = 1$ m and the battery delivers 1.5 V, the electric field has the strength of 1.5 V/m. The field acts on free electrons and sets them in motion against the direction of the field. Thus, the electric current starts flowing through the material. We can imagine a cross-section of the material through which passes electric charge q. The rate of the electric charge flowing (unit of charge per unit of time) is called electric current

$$i = \frac{dq}{dt} \tag{3.50}$$

The SI unit of current is ampere (A): 1A = 1 coulomb/1 s. In SI, ampere is defined as electric current, which is maintained in two infinitely long parallel wires separated by 1 m in free space, which produce a force between the two wires (due to their magnetic field) of 2×10^{-7} newtons for each meter of length. An ampere is quite strong electric current. In sensor technologies, generally much smaller currents are used, therefore, submultiples of A are often employed:
1 milliampere (mA)10^{-3} A

1 microampere (μA)	$= 10^{-6}$ A
1 nanoampere (nA)	$= 10^{-9}$ A
1 picoampere (pA)	$= 10^{-12}$ A
1 femtoampere (fA)	$= 10^{-15}$ A

No matter what the cross-section of the material is, whether it is homogeneous or not, the electric current through any cross-section is always the same for a given electric field. It is similar to water flow through a combination of serially connected pipes of different diameters: the rate of flow is the same throughout of the pipe combination. The water flows faster in the narrow sections and slower in the wide

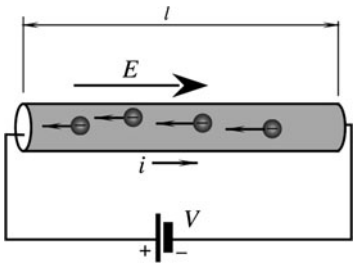

Fig. 3.16 Voltage across a material sets electric current

section, but amount of water passing through any cross-section per unit of time is constant. The reason for that is very simple: water in the pipes is neither drained out nor created. The same reason applies to electric current. One of the fundamental laws of physics is the law of conservation of charge. Under steady state conditions, charge in a material is neither created nor destroyed. *Whatever comes in must go out*. In this section, we do not consider any charge storages (capacitors), and all materials we discuss are said have pure resistive properties.

The mechanism of electrical conduction in a simplified form may be described as follows. A conducting material, say copper wire, can be modeled as a semirigid spring-like periodic lattice of positive copper ions. They are coupled together by strong electromagnetic forces. Each copper atom has one conduction electron, which is free to move about the lattice. When electric field **E** is established within the conductor, force $-e\mathbf{E}$ acts on each electron (e is the electron charge). The electron accelerates under the force and moves. However, the movement is very short as the electron collides with the neighboring copper atoms, which constantly vibrate with intensity, which is determined by the material temperature. The electron transfers its kinetic energy to the lattice and is often captured by the positive ion. When captured, it frees another electron, which keeps moving in the electric field until, in turn, it collides with the next portion of the lattice. The average time between collisions is designated as τ. It depends on the material type, structure, and impurities. For instance, at room temperature, a conduction electron in pure copper moves between collisions for an average distance of 0.04 μm with $\tau = 2.5 \times 10^{-14}$ s. In effect, electrons, which flow into the material near the negative side of the battery, are not the same that outflow to the positive terminal. However, the constant drift or flow of electrons is maintained throughout the material. Collisions of electrons with the material atoms further add to the atomic agitation and, subsequently, raise the material temperature. This is why passing of electric current through a resistive material results in the so called Joule heat liberation.

It was arbitrarily decided to define the direction of current flow along with the direction of the electric field, i.e., in the opposite direction of the electronic flow. Hence, the electric current flows from the positive to negative terminal of the battery while electrons actually move in the opposite direction. It is interesting to note that unlike water flowing through a pipe, electrons do not need to initially "fill up" the conductor before they start flowing out at a positive side. Electrons are

always present in a conductor. Since electric filed in a conductor propagates with the speed of light in the conductor's material, electric current appears at all parts of the conductor nearly instantaneously.

3.5.1 Specific Resistivity

If we fabricate two geometrically identical rods from different materials, say from copper and glass and apply to them the same voltage, the resulting currents will be quite different. A material may be characterized by its ability to pass electric current. It is called *resistivity* and material is said has electrical *resistance*, which is defined by Ohm's law, which means that a ratio of voltage to current is a constant

$$R = \frac{V}{i} \tag{3.51}$$

For the pure resistance (no inductance or capacitance) voltage and current are in-phase with each other, meaning that they are changing simultaneously.

Any material has electric resistivity[4] and therefore is called a *resistor*. The SI unit of resistance is 1 ohm (Ω) = 1 volt/1 ampere. Other multiples and submultiples of Ω are as follows:

1 milliohm (mΩ)	$= 10^{-3}$ Ω
1 kilohm (kΩ)	$= 10^{3}$ Ω
1 megohm (MΩ)	$= 10^{6}$ Ω
1 gigohm (GΩ)	$= 10^{9}$ Ω
1 terohm (TΩ)	$= 10^{12}$ Ω

If we compare electric current with water flow, pressure across the pipe line (Pascal) is analogous of voltage (V) across the resistor, electric current (C/s) is analogous of water flow (L/s) and electric resistance (Ω) corresponds to water flow resistance in the pipe (no special unit). It is clear that the resistance to water flow is smaller when the pipe is short, wide and has no obstructions. When the pipe has, for instance, a filter installed in it, resistance to water flow will be higher. Consider a human body where coronary blood flow may be restricted by cholesterol deposits on the inner lining of arteries. These deposits increase the flow resistance (called vascular resistance). The arterial blood pressure increases to compensate for rise in the vascular resistance but not always can keep up with it. If it is the case, the heart action that develops the arterial pressure is no longer sufficient to provide a necessary blood supply to vital organs, including a heart. This may result in a heart attack.

[4]Excluding superconductors, which are beyond the scope of this book.

The basic laws that govern the electric circuit designs are called Kirchhoff's Laws, after the German physicist Gustav Robert Kirchhoff (1824–1887). These laws were originally devised for the plumbing networks, which, as we have seen, are analogous to the electric networks.

Resistance is a characteristic of a device. It depends on both the material and the geometry of the resistor. Material itself can be characterized by a *specific resistivity*, ρ, which is defined as

$$\rho = \frac{E}{j}, \tag{3.52}$$

where current density $j = i/a$ (a is the area of the material cross section). The SI unit of resistivity is Ω·m. Resistivities of some materials are given in Appendix (Table A.7). Quite often, a reciprocal quantity is used which is called conductivity: $\sigma = 1/\rho$. The SI unit of conductivity is siemens having dimension [1/ω].

Resistivity of a material can be expressed through mean time between collisions, τ, the electronic charge, e, the mass of electron, m, and a number of conduction electrons per unit volume, n

$$\rho = \frac{m}{ne^2\tau}. \tag{3.53}$$

To find the resistance of a conductor the following formula may be used:

$$R = \rho\frac{l}{a}, \tag{3.54}$$

where a is the cross-sectional area and l is the length of the conductor. The ratio l/a is called a geometry factor.

Formula (3.54) establishes the fundamental relationship between resistance and its parameters. Thus, if one wants to design a resistive sensor, she should find ways of modulating either the specific resistivity or the geometry factor of the resistor.

3.5.2 Temperature Sensitivity

In reality, specific resistivity or conductivity of a material is not constant. It changes somewhat with temperature, t, and in a relatively narrow range may be linearly approximated through α, which is the temperature coefficient of resistance (TCR):

$$\rho = \rho_0\left(1 + \alpha\frac{t - t_0}{t_0}\right) \tag{3.55}$$

where ρ_0 is the specific resistivity at reference temperature t_0 (commonly either 0 or 25°C). In a broader range, resistivity is a nonlinear function of temperature.

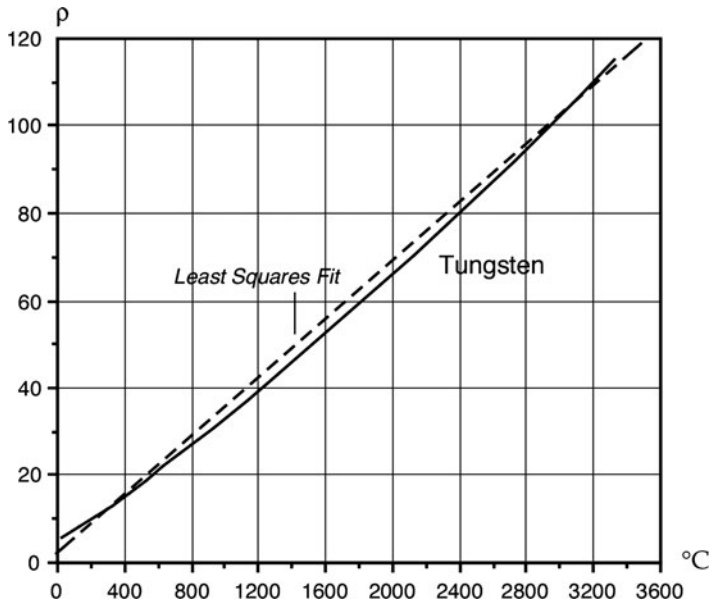

Fig. 3.17 Specific resistivity of tungsten as function of temperature

For nonprecision applications over a broad temperature range, resistivity of tungsten, as shown in Fig. 3.17 may be modeled by a best fit straight line. When better accuracy is required, the linear equation (3.55) should not be employed. Instead, higher order polynomials may be useful to model the resistivity. For instance, over a broader temperature range, tungsten resistivity may be found from the second order equation

$$\rho = \ 4.45 \ + 0.0269t + \ 1.914 \times 10^{-6} t^2, \tag{3.56}$$

where t is temperature in °C and ρ is in Ω·m.

Metals have positive temperature coefficients[5] (PTC) α, while many semiconductors and oxides have negative temperature coefficients of resistance (NTC).

When a resistor is used in an electronic circuit, its resistance shall be as temperature independent as possible. A "good" resistor may have $\alpha = 10^{-5}$ or even lower. However, in the sensor technologies, it is often desirable to have a "bad" resistor whose temperature coefficient of resistivity α is high and predictable. A strong α allows to fabricate a temperature sensor known as a thermistor;

[5]Since resistance of a metal increases with temperature, a tungsten filament in a light bulb acts as a self-regulator of temperature, so the filament does not burn out. When temperature increases, the resistance goes up and the current drops, causing the temperature to come down. If α for metals were negative, the filaments would instantly burn out and we would not have electric lights.

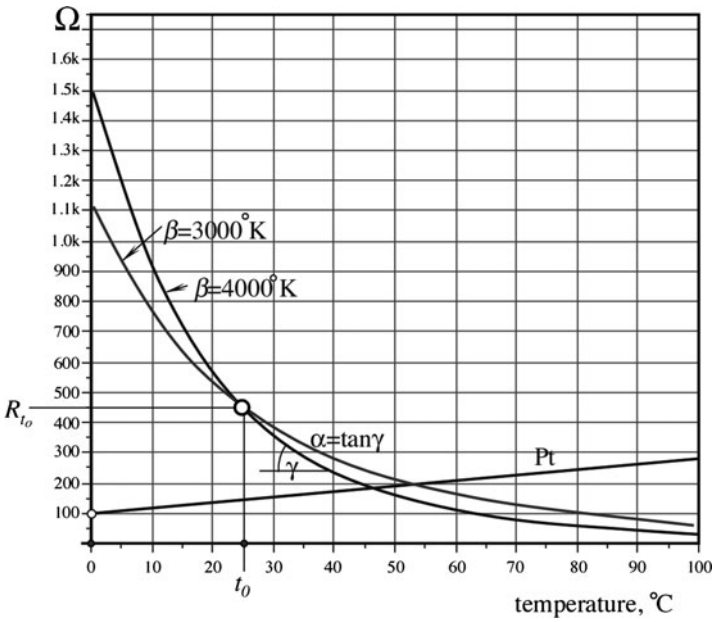

Fig. 3.18 Resistance-temperature characteristics for two thermistors and Pt RTD (Ro=1k); thermistors are calibrated at t_o=25°C and RTD at 0°C

(a contraction of words thermal and resistor) and the resistive temperature detector (RTD).[6] The most popular RTD is a platinum (Pt) temperature sensor, which operates over a broad temperature range from about −200°C to over 600°C. Resistance of Pt RTD is shown in Fig. 3.18. For a calibrating resistance R_0 at 0°C, the best fit straight line is given by equation

$$R = R_0(1 + 36.79 \times 10^{-4}t), \tag{3.57}$$

where t is temperature in °C and R is in Ω. The multiple at temperature (t) is the sensor's sensitivity (a slope), which may be expressed as +0.3679%/°C. A slight nonlinearity of the Pt resistance curve, if not corrected, may lead to an appreciable error over a broad temperature range. A better approximation of the Pt resistance is a second order polynomial which gives accuracy better than 0.01°C

$$R = R_0(1 + 39.08 \times 10^{-4}t - 5.8 \times 10^{-7}t^2)\Omega \tag{3.58}$$

It should be noted, however, that the coefficients in (3.57) and (3.58) somewhat depend on the material purity and manufacturing technologies. To compare

[6]See Section 16.3.1.

accuracies of the linear and second order models of the platinum thermometer, consider the following example. If a Pt RTD sensor at a reference temperature 0°C has resistivity R_0=100 Ω, at +150°C the linear approximation gives

$$R = 100 \cdot \left(1.0036 + 36.79 \times 10^{-4} \cdot 150\right) = 155.55\Omega,$$

while from the second-order approximation (3.58)

$$R = 100 \cdot \left(1 + 39.08 \times 10^{-4} \cdot 150 - 5.8 \times 10^{-7} \cdot 150^2\right) = 157.32\ \Omega.$$

The difference between the two is 1.76 Ω. This is equivalent to an error at +150°C of approximately −4.8°C (nearly 3%).

Thermistors are resistors with large either negative (NTC) or positive (PTC) temperature coefficients. The thermistors are ceramic semiconductors commonly made of oxides of one or more of the following metals: nickel, manganese, cobalt, titanium, iron. Oxides of other metals are occasionally used. Resistances of thermistors vary from a fraction of an ohm to many megohms. Thermistors can be produced in form of disks, droplets, tubes, flakes, or thin films deposited on ceramic substrates. Recent progress in thick film technology allows us to print thermistor on ceramic substrates. Resistance of a semiconductor may be controlled to create either NTC or PTC to form a semiconductive RTD.

Thermistors possess nonlinear temperature-resistance characteristics (Fig. 3.18), which are generally approximated by one of several different equations which are covered in detail in Chap. 16. The most popular of the thermistor's transfer function approximations is the exponential form

$$R_t = R_0 \mathrm{e}^{\beta\left(\frac{1}{T}-\frac{1}{T_0}\right)} \tag{3.59}$$

where T is the thermistor temperature, T_0 is the calibrating temperature, R_0 is the resistance at calibrating temperature T_0, and β is the material's characteristic temperature. All temperatures and β are in kelvin. Commonly, β ranges between 2,600 and 4,200 K and for a relatively narrow temperature range it can be considered temperature independent, which makes (3.59) a reasonably good approximation. When higher accuracy is required, other approximations are employed. Figure 3.18 shows resistance/temperature dependence of thermistors having β = 3,000 and 4,000°K and that for the platinum RTD. The platinum temperature sensor is substantially less sensitive and more linear with a positive slope, while thermistors are strongly nonlinear with a high sensitivity and negative slope.

Traditionally, thermistors are specified at temperature of t_0 = 25°C (T_0 = 298.15°K), while RTDs are specified at t_0 = 0°C (T_0 = 273.15°K).

3.5.3 Strain Sensitivity

Electrical resistance changes when the material is mechanically deformed. A mechanical deformation modulates either the specific resistivity or the geometry factor. The strain sensitivity is called the *piezoresistive effect*. As we have seen before, a "good" resistor better be stable, while having a "bad" resistor gives us an opportunity to make a sensor. In this case, we are talking about a strain sensor, which is the basis of many force and pressure sensors. The applied stress, σ, relates to force as

$$\sigma = \frac{F}{a} = E\frac{dl}{l} \tag{3.60}$$

where E is Young's modulus of the material, F is the applied force and a is the cross-sectional area. In this equation, the ratio $dl/l = e$ is called *strain*, which is a normalized deformation of the material.

Figure 3.19 shows a cylindrical conductor (wire) which is stretched by applied force F. Volume v of the material stays constant, while the length increases and the cross sectional area becomes smaller. As a result, (3.54) can be rewritten as

$$R = \frac{\rho}{v} l^2. \tag{3.61}$$

After differentiating, we can define sensitivity of resistance with respect to wire elongation

$$\frac{dR}{dl} = 2\frac{\rho}{v} l, \tag{3.62}$$

It follows from this equation that the sensitivity becomes higher for the longer and thinner wires with high specific resistance. Normalized incremental resistance of the strained wire is a linear function of strain, e, and it can be expressed as

$$\frac{dR}{R} = S_e e, \tag{3.63}$$

where S_e is known as the *gauge factor* or *sensitivity* of the strain gauge element. For metallic wires it ranges from 2 to 6. It is much higher for the semiconductor gauges

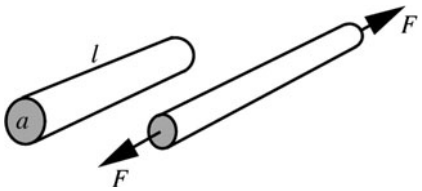

Fig. 3.19 Strain changes geometry of a conductor and its resistance

where it is between 40 and 200, because in the semiconductors, the geometry factor plays a much smaller role than change in the specific resistivity due to deformation of the crystalline structure of the material.

Early strain gauges where metal filaments. The gauge elements were formed on a backing film of electrically isolating material. Today, they are manufactured form constantan (copper/nickel alloy) foil or single crystal semiconductor materials (silicon with boron impurities). The gauge pattern is formed either by mechanical cutting or photochemical etching. When a semiconductor material is stressed, its resistivity changes depending on the type of the material and the doping dose (see Sect. 9.1). However, the strain sensitivity in semiconductors is temperature dependent which requires a proper compensation when used over a broad temperature range.

3.5.4 Moisture Sensitivity

By selecting material for a resistor, one can control its specific resistivity and susceptibility of such to the environmental factors. One of the factors that may greatly affect ρ is the amount of moisture that can be absorbed by the resistor. A moisture-dependent resistor can be fabricated of a hygroscopic material whose specific resistivity is strongly influenced by concentration of the absorbed water molecules. This is the basis for the resistive humidity sensors, which are called hygristors.

A typical hygristor is comprised of a ceramic substrate that has two silk-screen printed conductive interdigitized [7] electrodes (Fig. 3.20a). The electrodes are metal conductors. The electrodes and the space between them are covered by hygroscopic

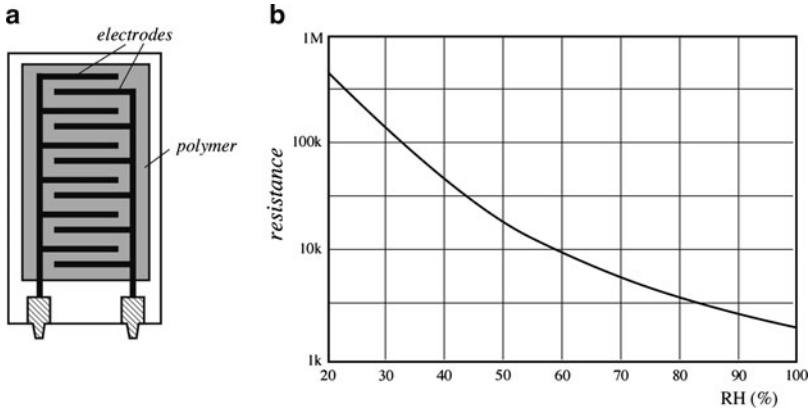

Fig. 3.20 Hygristor design (**a**) and its transfer function (**b**)

[7]The term is based on a similarity between the electrode shape and fingers (digits) of two human hands grasping one another.

semiconductive gel which forms a matrix to hold the conductive particles. As a result, a resistor is formed between two electrodes. The gel [3] is typically fabricated of hydroxyethylcellulose, nonyl phenyl polyethylene glycol ether (a nonchemist needs to exercise her tongue to say the name!), and other organic materials with addition of carbon powder. The gel is thoroughly milled to produce a smooth mixture. Another type of a hygristor is fabricated of lithium chloride (LiCl) film and a binder. The sensor substrates are dipped into the milled gel at controlled rates. The coated substrates are cured under controlled temperature and humidity. Resistance of the coating changes with humidity in a nonlinear way (Fig. 3.20b), which can be taken into account during calibration and data processing. The response time for most hygristors ranges from 10 to 30 s. The resistance range varies from 1 kΩ to 100 MΩ.

The hygristors are active sensors, that is, they require an excitation signal to produce an electrical output. It is important to use only symmetrical AC excitation current with no DC bias to prevent polarization of the coating or the sensor will be destroyed.

3.6 Piezoelectric Effect

The piezoelectric effect is generation of electric charge by a crystalline material upon subjecting it to stress. The effect exists in natural crystals, such as quartz (chemical formula SiO_2), and poled (artificially polarized) human-made ceramics and some polymers, such as PVDF. It is said that piezoelectric material possesses ferroelectric properties. The name was given by an analogy with ferromagnetic properties, though there is no iron in most piezoelectrics. The word *piezo* comes from the Greek πιεση meaning "to press." The Curie brothers discovered the piezoelectric effect in quartz in 1880, but very little practical use was made until 1917 when another Frenchman, professor P. Langevin used x-cut plates of quartz to generate and detect sound waves in water. His work led to the development of sonar.

A simplified, yet quite explanatory model of the piezoelectric effect was proposed in 1927 by A. Meissner [4]. A quartz crystal is modeled as a helix (Fig. 3.21a)

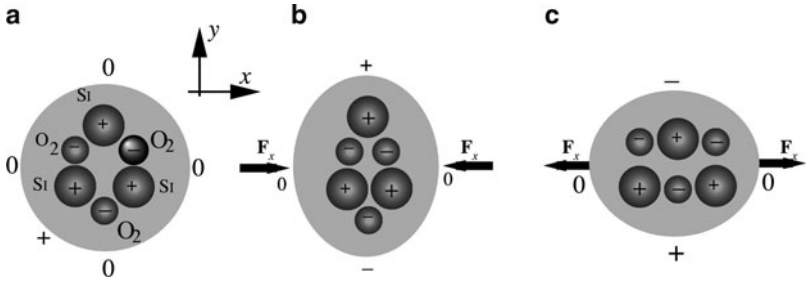

Fig. 3.21 Piezoelectric effect in a quartz crystal

with one silicon, Si, and two oxygen, O_2, atoms alternating around the helix. A quartz crystal is cut along its axes x, y, and z, thus Fig. 3.21a is a view along the z-axis. In a single crystal-cell there are three atoms of silicon and six oxygen atoms. Oxygen is being lumped in pairs. Each silicon atom carries four positive charges and a pair of oxygen atoms carries four negative charges (two per atom). Therefore a quartz cell is electrically neutral under the no-stress conditions. When external force, F_x, is applied along the x-axis, the hexagonal lattice becomes deformed. Figure 3.21b shows a compressing force, which shifts atoms in a crystal in such a manner as a positive charge is built up at the silicon atom side and the negative at the oxygen pair side. Thus, the crystal develops an electric charge along the y-axis. If the crystal is stretched along the x-axis (Fig. 3.21c), a charge of opposite polarity is built along the y-axis, which is a result of a different deformation. This simple model illustrates that crystalline material can develop electric charge on its surface in response to a mechanical deformation. A similar explanation may be applied to the pyroelectric effect, which is covered in the next section of this chapter.

To pickup an electric charge, conductive electrodes must be applied to the crystal at the opposite sides of the cut (Fig. 3.22). As a result, a piezoelectric sensor becomes a capacitor with a dielectric material between the metal plates, where the dielectric is a piezoelectric crystalline material. The dielectric acts as a generator of electric charge, resulting in voltage V across the capacitor. Although charge in a crystalline dielectric is formed at the location of an acting force, metal electrodes equalize charges along the surface making the capacitor not selectively sensitive. However, if electrodes are formed with a complex pattern (e.g., multiple electrodes), it is possible to determine the exact location of the applied force by measuring the response from a selected electrode.

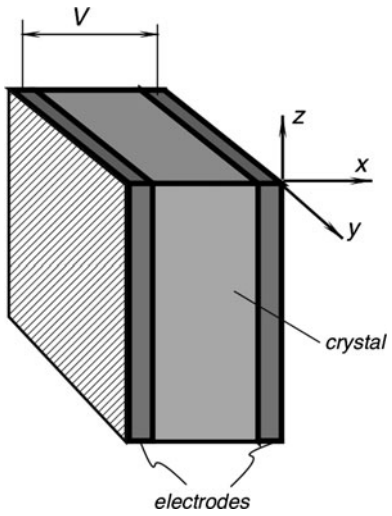

Fig. 3.22 Piezoelectric sensor is formed by applying electrodes to a poled crystalline material

The piezoelectric effect is a reversible physical phenomenon. This means that applying voltage across the crystal produces mechanical strain. It is possible by placing several electrodes on the crystal to use one pair of electrodes to deliver voltage to the crystal and the other pair of electrodes to pick up charge resulting from developed strain. This method is used quite extensively in various piezoelectric transducers.

The magnitude of the piezoelectric effect in a simplified form can be represented by the vector of polarization [5]

$$\mathbf{P} = \mathbf{P}_{xx} + \mathbf{P}_{yy} + \mathbf{P}_{zz}, \tag{3.64}$$

where x, y, and z refer to a conventional orthogonal system related to the crystal axes. In terms of axial stress, $\boldsymbol{\sigma}$, we can write[8]

$$\begin{aligned} \mathbf{P}_{xx} &= d_{11}\boldsymbol{\sigma}_{xx} + d_{12}\boldsymbol{\sigma}_{yy} + d_{13}\boldsymbol{\sigma}_{zz}, \\ \mathbf{P}_{\mathrm{yy}} &= d_{21}\boldsymbol{\sigma}_{xx} + d_{22}\boldsymbol{\sigma}_{yy} + d_{23}\boldsymbol{\sigma}_{zz}, \\ \mathbf{P}_{\mathrm{zz}} &= d_{31}\boldsymbol{\sigma}_{xx} + d_{32}\boldsymbol{\sigma}_{yy} + d_{33}\boldsymbol{\sigma}_{zz}, \end{aligned} \tag{3.65}$$

where constants d_{mn} are the piezoelectric coefficients along the orthogonal axes of crystal cut. Dimensions of these coefficients are C/N (coulomb/newton), that is, charge unit per unit force.

For the convenience of computation, two additional units have been introduced. The first is a g-coefficient, which is defined by a division of corresponding d_{mn}-coefficients by the absolute dielectric constant

$$g_{mn} = \frac{d_{mn}}{\varepsilon_0 \varepsilon_{mn}}. \tag{3.66}$$

This coefficient represents a voltage gradient (electric field) generated by the crystal per unit applied pressure, i.e., its dimension is

$$\frac{V}{m} \Big/ \frac{V}{m^2}$$

Another coefficient, which is designated h, is obtained by multiplying the g-coefficients by the corresponding Young's moduli for the corresponding crystal axes. Dimension of the h-coefficient is

$$\frac{V}{m} \Big/ \frac{m}{m}$$

[8]The complete set of coefficients also includes shear stress and the corresponding d-coefficients.

Piezoelectric crystals are direct converters of mechanical energy into electrical. The efficiency of the conversion can be determined from the so-called *coupling coefficients kmn*:

$$k_{mn} = \sqrt{d_{mn} h_{mn}} \tag{3.67}$$

The k-coefficient is an important characteristic for applications where energy efficiency is of a prime importance, like in the acoustic and ultrasonic sensors.

Charge generated by the piezoelectric crystal is proportional to applied force, for instance, in the x-direction the charge is

$$Q_x = d_{11} F_x. \tag{3.68}$$

Since a crystal with deposited electrodes forms a capacitor having capacitance, C, voltage, V, which develops across between the electrodes is

$$V = \frac{Q_x}{C} = \frac{d_{11}}{C} F_x \tag{3.69}$$

In turn, the capacitance can be represented [see (3.23)] through the electrode surface area,[9] a, and the crystal thickness, l:

$$C = \kappa \varepsilon_0 \frac{a}{l}. \tag{3.70}$$

Then, the output voltage is

$$V = \frac{d_{11}}{C} F_x = \frac{d_{11} l}{\kappa \varepsilon_o a} F_x. \tag{3.71}$$

3.6.1 Ceramic Piezoelectric Materials

The manufacturing of ceramic PZT sensors begins with high-purity metal oxides (lead oxide, zirconium oxide, titanium oxide, etc.) in the form of fine powders having various colors. The powders are milled to a specific fineness, and mixed thoroughly in chemically correct proportions. In a process called "calcining," the mixtures are then exposed to an elevated temperature, allowing the ingredients to react to form a powder, each grain of which has a chemical composition close to the

[9]The electrode, not the crystal area! Piezo induced charge can be collected only over the area covered by the electrode.

desired final composition. At this stage, however, the grain does not have yet the desired crystalline structure.

The next step is to mix the calcined powder with solid and/or liquid organic binders (intended to burn out during firing) and mechanically form the mixture into a "cake," which closely approximates a shape of the final sensing element. To form the "cakes" of desired shapes, several methods can be used. Among them are pressing (under force of a hydraulic powered piston), casting (pouring viscous liquid into molds and allowing to dry), extrusion (pressing the mixture through a die, or a pair of rolls to form thin sheets), and tape casting (pulling viscous liquid onto a smooth moving belt).

After the "cakes" have been formed, they are placed into a kiln and exposed to a very carefully controlled temperature profile. After burning out of organic binders, the material shrinks by about 15%. The "cakes" are heated to a red glow and maintained at that state for some time, which is called the "soak time," during which the final chemical reaction occurs. The crystalline structure is formed when the material is cooled down. Depending on the material, the entire firing may take 24 h.

When the material is cold, the contact electrodes are applied to its surface. This can be done by several methods. The most common of them are: a fired-on silver (a silk-screening of silver-glass mixture and re-firing), an electroless plating (a chemical deposition in a special bath), and a sputtering (an exposure to metal vapor in a partial vacuum).

Crystallities (crystal cells) in the material can be considered electric dipoles. In some materials, like quartz, these cells are naturally oriented along the crystal axes, thus giving the material sensitivity to stress. In other materials, the dipoles are randomly oriented and the materials need to be "poled" to possess piezoelectric properties. To give a crystalline material piezoelectric properties, several poling techniques can be used. The most popular poling process is a thermal poling, which includes the following steps:

1. A crystalline material (ceramic or polymer film), which has randomly oriented dipoles (Fig. 3.23a), is warmed up slightly below its Curie temperature. In some cases (e.g., for a PVDF film) the material is stressed. High temperature results in stronger agitation of dipoles and permits to orient them more easily in a desirable direction.

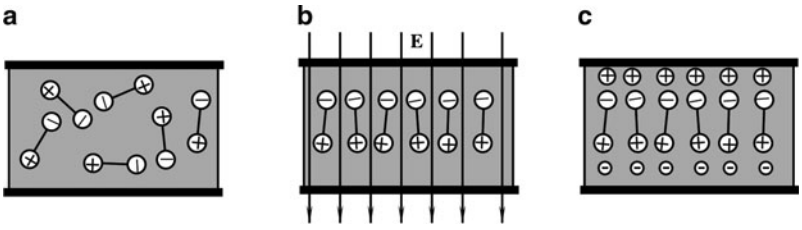

Fig. 3.23 Thermal poling of a piezo- and pyroelectric material

2. Material is placed in strong electric field, **E** (Fig. 3.23b) where dipoles align along the field lines. The alignment is not total. Many dipoles deviate from the filed direction quite strongly, however, statistically predominant orientation of the dipoles is maintained.
3. The material is cooled down while the electric field across its thickness is maintained.
4. The electric field is removed and the poling process is complete. As long as the poled material is maintained below the Curie temperature, its polarization remains permanent. The dipoles stay "frozen" in the direction, which was given to them by the electric field at high temperature (Fig. 3.23c).

Another method called a corona discharge poling. It''s used to produce polymer piezo/pyroelectric films (see Sect. 3.6.2). The film is subjected to a corona discharge from an electrode at several million volts per cm of film thickness for 40–50 s [6, 7]. Corona polarization is uncomplicated to perform and can be easily applied before electric breakdown occurs, making this process useful at room temperature.

The final operation in preparation of the sensing element for shaping and finishing. This includes cutting, machining, and grinding. After the piezo (pyro) element is prepared, it is installed into a sensor's housing, where its electrodes are bonded to electrical terminals and other electronic components.

After poling, the crystal remains permanently polarized, with an electric charge formed at the electrodes for a relatively short time. There is a sufficient amount of free charge carriers, which move in the electric field setup inside the bulk material and there are plenty charged ions in the surrounding air. The charge carriers move toward the poled dipoles and neutralize their charges (see Fig. 3.23c). Hence, after a while, the poled piezoelectric material becomes electrically discharged as long as it remains under steady-state conditions. When stress is applied, or air blows near its surface (Sect. 10.7) the balanced state is degraded and the piezoelectric material develops an electric charge. If the stress is maintained, the charges again will be neutralized by the internal leakage. Thus, a piezoelectric sensor is responsive only to a changing stress rather than to a steady level of it. In other words, a piezoelectric sensor is an AC device, rather than a DC device.

Piezoelectric directional sensitivities (d coefficients) are temperature-dependent. For some materials (quartz), sensitivity drops with a slope of $-0.016\%/^{\circ}C$. For others (the PVDF films and ceramics) at temperatures below 40°C, it may drop and at higher temperatures it increases with a raise in temperature. Nowadays, the most popular materials for fabrication of piezoelectric sensors are ceramics [8–10]. The earliest of the ferroelectric ceramics was barium titanate, a polycrystalline substance having the chemical formula $BaTiO_3$. The stability of permanent polarization relies on the coercive force of the dipoles. In some materials, polarization may decrease with time. To improve stability of poled material, impurities have been introduced in the basic material with the idea that the polarization may be "locked" into position [5]. While the piezoelectric constant changes with operating temperature, a dielectric constant, ε, exhibits a similar dependence. Thus, according to

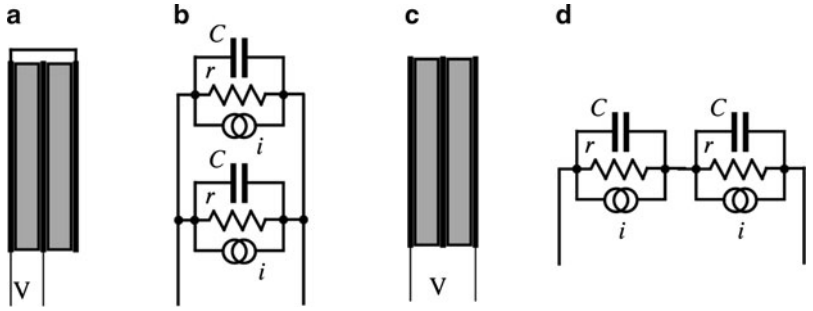

Fig. 3.24 Parallel (**a**) and serial (**c**) laminated piezoelectric sensors and their corresponding equivalent circuits (**b**), (**d**)

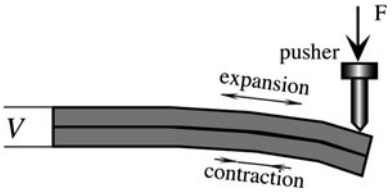

Fig. 3.25 Laminated two-layer piezoelectric sensor

(3.71), variations in these values tend to cancel each other as they are entered into numerator and denominator. This results in a better stability of the output voltage, V, over a broad temperature range.

The piezoelectric elements may be used as a single crystal, or in a multilayer form where several plates of the material are laminated together. This must be done with electrodes placed in-between. Figure 3.25 shows a two-layer force sensor.[10] When an external force is applied, the upper part of the sensor expands while the bottom compresses. If the layers are laminated correctly, this produces a double output signal. Double sensors can have either a parallel connection as shown in Fig. 3.24a, or a serial connection as in Fig. 3.24c. The electrical equivalent circuit of the piezoelectric sensor is a parallel connection of a stress-induced current source (i), leakage resistance (r), and capacitance (C). Depending on the layer connection equivalent circuits for the laminated sensors are as shown in Figs. 3.24b, d. The leakage resistors r are very large, on the orders of 10^{12}–10^{14} Ω, which means that the sensor has an extremely high output impedance. This requires special interface circuits, such as charge and current-to-voltage converters, or voltage amplifiers with high input resistances.

Since silicon does not possess piezoelectric properties, such properties can be added on by depositing crystalline layers of the piezoelectric materials. The three most popular materials are zinc oxide (ZnO), aluminum nitride (AlN), and the

[10]Remember, a piezoelectric sensor is an AC device, so it will not respond to a constant or slowly changing force.

so-called solid solution system of lead-zirconite-titanium oxides $Pb(Zr,Ti)O_3$ known as PZT ceramic, basically the same material used for fabrication of discrete piezoelectric sensors as described above.

Zinc oxide in addition to the piezoelectric properties also is pyroelectric. It was the first and most popular material for development of ultrasonic acoustic sensors, surface acoustic wave (SAW) devices, microbalances, etc. One of its advantages is the ease of chemical etching. The zinc oxide thin films are usually deposited on silicon by employing the sputtering technology.

Aluminum nitride (AlN) is an excellent piezoelectric material because of its high acoustic velocity and its endurance in humidity and high temperature. Its piezoelectric coefficient is somewhat lower than in ZnO but higher than in other thin-film piezoelectric materials, excluding ceramics. The high acoustic velocity makes it an attractive choice in the GHz frequency range. Usually, the AlN thin films are fabricated by using the chemical vapor deposition (CVD) or reactive molecular beam epitaxy (MBE) technologies. However, the drawback of using these deposition methods is the need for high heating temperature (up to 1,300°C) of the substrate.

The PZT thin films possess a larger piezoelectric coefficient than ZnO or AlN, and also a high pyroelectric coefficient, which makes it a good candidate for fabrication of the thermal radiation detectors. A great variety of deposition techniques is available for the PZT, among which are the electron-beam evaporation [11], RF sputtering [12], ion-beam deposition [13], epitaxial growth by RF sputtering [14], magnetron sputtering [15], laser ablation [16], and sol-gel [17].

3.6.2 Polymer Piezoelectric Films

In 1969, H. Kawai discovered a strong piezoelectricity in polyvinylidene fluoride (PVDF) and in 1975 the Japanese company Pioneer, Ltd. developed the first commercial product with the PVDF as a piezoelectric loudspeakers and earphones [18]. PVDF is a semicrystalline polymer with an approximate degree of crystallinity of 50% [19]. Like other semicrystalline polymers, PVDF consists of a lamellar structure mixed with amorphous regions. The chemical structure of it contains the repeat unit of doubly fluorinated ethene CF_2-CH_2:

$$\left[\begin{array}{cc} H & F \\ | & | \\ -C- & -C- \\ | & | \\ H & F \end{array}\right]_n$$

PVDF molecular weight is about 10^5, which corresponds to about 2,000 repeat units. The film is quite transparent in the visible and near-IR region, and is absorptive in the mid- and far-infrared portions of the electromagnetic spectrum.

The polymer melts near 170°C. Its density is about 1,780 kg/m^3. PVDF is a mechanically durable and flexible material. In piezoelectric applications, it is usually drawn, uniaxially or biaxially to several times its length. Elastic constants, for example, Young modulus, depend on this draw ratio. Thus, if the PVDF film was drawn at 140°C to the ratio of 4:1, the modulus value is 2.1 GPa, while for the draw ratio of 6.8:1 it was 4.1 GPa. Resistivity of the film also depends on the stretch ratio. For instance, at low stretch it is about 6.3×10^{15} Ωcm, while for the stretch ratio 7:1 it is 2×10^{16} Ωcm.

PVDF does not have a higher, or even as high piezoelectric coefficient as other commonly used materials, like $BaTiO_3$ or PZT. However, it has a unique quality not to depolarize while being subjected to very high alternating electric fields. This means that even though the value of d_{31} of PVDF is about 10% of PZT, the maximum strain observable in PVDF will be 10 times larger than in PZT since the maximum permissible field is 100 times greater for PVDF. The film exhibits good stability: when stored at 60°C it loses its sensitivity by about 1–2% over 6 months. Comparative characteristics for various piezoelectric materials are given in Table A.8. Another advantage of piezo film over piezo ceramic is its low acoustic impedance which is closer to that of water, human tissue and other organic materials. For example, the acoustic impedance of piezo film is only 2.6 times that of water, whereas piezo ceramics are typically 11 times greater. A close impedance match permits more efficient transduction of acoustic signals in water and tissue.

Some unique properties of the piezoelectric films are (from Measurement Specialties, Inc., www.msiusa.com)

- Wide frequency range - 0.001 Hz to 10^9 Hz.
- Vast dynamic range (10^{-8} to 10^6 psi or μtorr to Mbar).
- Low acoustic impedance – close match to water, human tissue and adhesive systems.
- High elastic compliance.
- High voltage output – 10 times higher than piezo ceramics for the same force input.
- High dielectric strength – withstanding strong fields (75 V/μm) where most piezo ceramics depolarize.
- High mechanical strength and impact resistance (10^9—10^{10} Pascal modulus).
- High stability—resisting moisture (<0.02% moisture absorption), most chemicals, oxidants, and intense ultraviolet and nuclear radiation.
- Can be fabricated into many shapes.
- Can be glued with commercial adhesives.

Typical properties of piezoelectric films are given in Table 3.1.

Like some other ferroelectric materials, PVDF is also pyroelectric (see Sect. 3.7), producing electrical charge in response to a change in temperature. PVDF strongly absorbs infrared energy in the 7–20 μm wavelengths, covering the same wavelength spectrum as heat from the human body. However, in spite that the film can absorb thermal radiation, a pyroelectric sensor has the film sandwiched between

Table 3.1 Typical properties of piezoelectric films (from [12])

Symbol	Parameter	PVDF	Copolymer	Units
t	Thickness	9, 28, 52, 110	<1 to 1200	μm (micron, 10^{-6})
d_{31}	Piezo strain constant	22	11	$10^{-12}\frac{m/m}{v/m}$ or $\frac{C/m^2}{N/m^2}$
d_{33}		−33	−38	
g_{31}	Piezo stress constant	216	162	$10^{-3}\frac{V/m}{N/m^2}$ or $\frac{m/m}{C/m^2}$
g_{33}		−330	−542	
k_{31}	Electromechanical coupling	12%	20%	
k_t	factor	14%	25–29%	
C	Capacitance	380 for 28 μm	68 for 100 μm	pF/cm^2 @ 1 kHz
Y	Young's modulus	2–4	3–5	10^9 N/m^2
V_0	Speed of Stretch:	1.5	2.3	10^3 m/s
	sound Thickness:	2.2	2.4	
p	Pyroelectric coefficient	30	40	10^{-6} C/m^2 °K
ε	Permittivity	106–113	65–75	10^{-12} F/m
$\varepsilon/\varepsilon_0$	Relative permittivity	12–13	7–8	
ρ_m	Mass density	1.78	1.82	10^3 kg/m
ρ_e	Volume resistivity	$>10^{13}$	$>10^{14}$	Ohm meters
$R_\square$	Surface metallization resistivity	<3.0	<3.0	Ohms/square for NiAl
$R_\square$		0.1	0.1	Ohms/square for Ag lnk
tan δ_e	Loss tangent	0.02	0.015	@ 1 kHz
	Yield strength	45–55	20–30	10^6 N/m^2 (stretch axis)
	Temperature Range	−40 to 80...100	−40 to 115...145	°C
	Water absorption	<0.02	<0.02	% H_2O
	Maximum operating voltage	750 (30)	750(30)	V/mil(V/μm), DC, @ 25°C
	Breakdown voltage	2,000 (80)	2,000 (80)	V/mil(V/μm), DC, @ 25°C

two thin metal electrodes, which can be quite reflective in the spectral range of interest, so no infrared radiation can penetrate the electrodes and be absorbed by the film. To resolve this difficulty, the electrode that is exposed to thermal radiation either is coated with heat absorbing layer or is made of nichrome, a metal alloy having high infrared absorptivity. When the thermal radiation is absorbed, it is converted into heat that quickly propagates though the PVDF film by means of thermal conduction (see Sect. 3.7).

The PVDF film makes a useful human motion sensor as well as pyroelectric sensor for more sophisticated applications like vidicon cameras for night vision and laser beam profiling sensors. A dense infrared array was introduced to identify one's fingerprint pattern using the pyroelectric effect of the polymer. The copolymers of PVDF have expanded the applications of piezoelectric polymer sensors. These copolymers permit use at higher temperatures (135°C) and offer desirable new sensor shapes, like cylinders and hemispheres. Thickness extremes are possible with copolymer that cannot be readily attained with PVDF. These include ultrathin

(200 Å) spin-cast coatings that enable new sensor-on-silicon applications, and cylinders with wall thicknesses in excess of 1,200 μm for sonar. Piezo cable is also produced using copolymer.

Unlike the piezoelectric ceramic transducers, the piezo film transducers offer wide dynamic range and are also broadband. These wide band characteristics (near DC to 2 GHz) and low Q are partly attributable to the softness of polymers. As audio transmitters, a curved piezo film element, clamped at each end, vibrates in the length (d_{31}) mode. The d_{31} configuration is also used for air ultrasound ranging applications up to frequencies of about 50 kHz. When used as a high ultrasonic transmitter (generally >500 kHz), the piezo film is normally operated in the thickness (d_{33}) mode. Maximum transmission occurs at thickness resonance. The basic half-wavelength resonance of 28 μm piezo film is about 40 MHz: Resonance values depend on film thickness. They range from low MHz for thick films to >100 MHz for very thin films (μm range).

When extruded into thin film, piezoelectric polymers can be directly attached to a structure without disturbing its mechanical motion. Piezo film is well suited to strain sensing applications requiring very wide bandwidth and high sensitivity. As an actuator, the polymer's low acoustic impedance permits the efficient transfer of a broadband of energy into air and other gases.

Piezo film does have some limitations for certain applications. It makes a relatively weak electromechanical transmitter when compared to ceramics, particularly at resonance and in low-frequency applications. The copolymer film has maximum operating/storage temperatures as high as 135°C, while PVDF is not recommended for use or storage above 100°C. Also, if the electrodes on the film are exposed, the sensor can be sensitive to electromagnetic radiation. Good shielding techniques are available for high EMI/RFI environments. Table A.8 provides a comparison of the piezoelectric properties of PVDF polymer and other popular piezoelectric ceramic materials.

Piezoelectric effect is the prime means of converting mechanical deformation into electrical signal and vice versa in the miniature semiconductor sensors. Hence, it allows to construct a direct sensor for converting mechanical energy into electric signal. As it was mentioned above, the effect can be used only for converting the changing stimuli and cannot be used for conversion of steady state or very slow changing signals.

3.7 Pyroelectric Effect

The pyroelectric materials are crystalline substances capable of generating an electrical charge in response to heat flow. The pyroelectric effect is very closely related to the piezoelectric effect. Before going further, we recommend that the readers familiarize themselves with Sect. 3.6.

Like piezoelectrics, the pyroelectric materials are used in form of thin slices or films with electrodes deposited on the opposite sides to collect the thermally

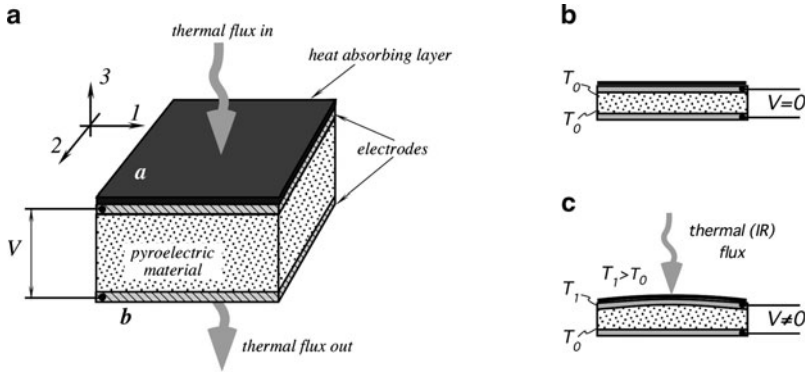

Fig. 3.26 Pyroelectric sensor has two electrodes at the opposite sides of the crystal (**a**). Thermal radiation is applied along axis 3 from the top and absorbed by the heat absorbing layer. Heat conductively travels through the pyroelectric material and is partially emanated downward from side *a*. Pyroelectric sensor in a neutral state (**b**); heat expands the upper layer, resulting in a piezoelectric charge (**c**)

induced charges (Fig. 3.26a). The pyroelectric sensor is essentially a capacitor, which can be electrically charged by flux of heat. The detector does not require any external electrical bias (excitation signal), thus it is a direct converter of heat into electricity. It needs only an appropriate electronic interface circuit to measure the charge. Contrary to thermoelectrics (thermocouples), which produce a steady voltage when two dissimilar metal junctions are held at steady but different temperatures (see Sect. 3.9), pyroelectrics generate charge in response to a *change* in temperature. Since a change in temperature essentially requires propagation of heat, a pyroelectric device is a heat flow detector rather than heat detector. When a pyroelectric crystal is exposed to a heat flow (for instance, from an infrared radiation source or from touching a warm object), temperature of the exposed side is elevates and the side becomes a source of heat, which propagates through the pyroelectric material toward its opposite side. Hence, there is an outflow of heat from the crystal to the environment, as it is shown in Fig. 3.26a.

A crystal is considered to be pyroelectric if it exhibits a spontaneous temperature-dependent polarization. Of the 32 crystal classes, 21 are noncentrosymmetric and ten of these exhibit pyroelectric properties. Beside pyroelectric properties, all these materials exhibit piezoelectric properties as well; they generate an electrical charge in response to mechanical stress. Thus, when a pyroelectric sensor is designed, it is very important to minimize all potential mechanical disturbances.

Pyroelectricity was observed for the first time in tourmaline crystals in the eighteenth century (some claim that the Greeks noticed it 23 centuries ago). Later, in the nineteenth century, Rochelle salt was used to make pyroelectric sensors. A large variety of materials became available after 1915: KDP (KH_2PO_4), ADP ($NH_4H_2PO_4$), $BaTiO_3$, and a composite of $PbTiO_3$ and $PbZrO_3$ known as PZT.

Presently, more than 1,000 materials with reversible polarization are known. They are called ferroelectric crystals.[11] The most important among them are triglycine sulfate (TGS) and lithium tantalate ($LiTaO_3$). In 1969 H. Kawai discovered strong piezoelectricity in the plastic materials, polyvinyl fluoride (PVF), and polyvinylidene fluoride (PVDF) [21]. These materials also possess substantial pyroelectric properties.

A pyroelectric material can be considered as a composition of a large number of minute crystallities, where each behaves as a small electric dipole. All these dipoles are randomly oriented (Fig. 3.23a). Above a certain temperature, known as the Curie point, the crystallities have no dipole moment. Manufacturing (poling) of pyroelectric materials is similar to that of the piezoelectrics (see Sect. 3.6).

There are several mechanisms by which changes in temperature will result in pyroelectricity. Temperature changes may cause shortening or elongation of individual dipoles. It may also affect the randomness of the dipole orientations due to thermal agitation. These phenomena are called primary pyroelectricity. There is also secondary pyroelectricity, which, in a simplified way, may be described as a result of the piezoelectric effect, that is, a development of strain in the material due to thermal expansion. Figure 3.26b shows a pyroelectric sensor whose temperature T_0 is homogeneous over its volume. That is, the sensor generates zero voltage across the electrodes. Now, let us assume that heat is applied to the top side of the sensor (Fig. 3.26c) in form of thermal (infrared) radiation. The radiation is absorbed by the heat absorbing layer (e.g., goldblack or organic paint) and warms up the upper side of the pyroelectric material. As a result of the heat absorption, the upper side becomes warmer (the new warmer temperature is T_1), which causes the top side of the material to expand. The expansion leads to flexing of the sensor which, in turn, produces stress and change in a dipole orientation. Being a piezoelectric, the stressed material generates electric charges of the opposite polarities on the electrodes and thus a voltage is observed across the electrodes. Hence, we may regard a secondary pyroelectricity as a sequence of events: thermal radiation-heat absorption-thermally induced stress-electric charge.

Let us analyze properties of a pyroelectric material. The dipole moment, M, of the bulk pyroelectric sensor is

$$M = \mu A h, \tag{3.72}$$

where μ is the dipole moment per unit volume, A is the sensor's area, and h is the thickness. The charge, Q_a, which can be picked up by the electrodes, develops the dipole moment across the material

$$M_0 = Q_a \cdot h. \tag{3.73}$$

[11]This is a misnomer as the prefix *ferro*, meaning *iron*, is used despite the fact that most ferroelectric materials do not have iron in their lattice. It is used by analogy with ferromagnetics.

M must be equal to M_0, so that

$$Q_a = \mu \cdot A \tag{3.74}$$

As the temperature varies, the dipole moment also changes, resulting in an induced charge. Thermal absorption may be related to a dipole change, so that μ must be considered as function of both temperature, T_a, and an incremental thermal energy, ΔW, absorbed by the material

$$\Delta Q_a = A \cdot \mu(T_a, \Delta W) \tag{3.75}$$

Figure 3.27 depicts a pyroelectric detector (pyroelectric sensor) connected to a resistor R_b that represents either the internal leakage resistance or a combined input resistance of the interface circuit, which is connected to the sensor. The equivalent electrical circuit of the sensor is shown at right. It consists of three components: (1) the current source generating a heat induced current, i (remember that a current is a movement of electric charges), (2) the sensor's capacitance, C, and (3) the leakage resistance, R_b.

The output signal from the pyroelectric sensor can be taken in form of either charge (current) or voltage, depending on the application. Being a capacitor, the pyroelectric device is discharged when connected to a resistor, R_b. Electric current through the resistor and voltage across the resistor represent the heat flow induced charge. It can be characterized by two pyroelectric coefficients [22]

$$P_Q = \frac{dP_s}{dT} \quad \text{Pyroelectric charge coefficient}$$

$$P_V = \frac{dE}{dT} \quad \text{Pyroelectric voltage coefficient} \tag{3.76}$$

where P_s is the spontaneous polarization (which is the other way to say "electric charge"), E is the electric field strength, and T is the temperature in K. Both coefficients are related by way of the electric permitivity, ε_r, and dielectric constant, ε_0

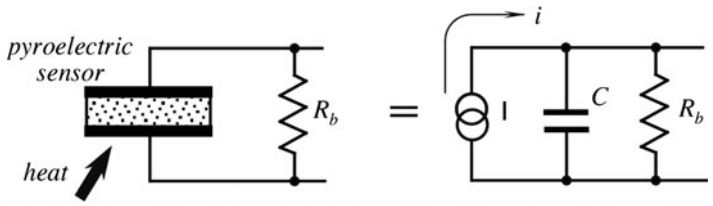

Fig. 3.27 Pyroelectric sensor and its equivalent circuit

$$\frac{P_Q}{P_V} = \frac{dP_s}{dE} = \varepsilon_r \cdot \varepsilon_0 \tag{3.77}$$

The polarization is temperature dependent and, as a result, both pyroelectric coefficients (3.76) are also functions of temperature.

If a pyroelectric material is exposed to a heat source, its temperature rises by ΔT and the corresponding charge and voltage changes can be described by the following equations:

$$\Delta Q = P_Q A \Delta T \tag{3.78}$$

$$\Delta V = P_V h \Delta T \tag{3.79}$$

Remembering that the sensor's capacitance can be defined as

$$C_e = \frac{\Delta Q}{\Delta V} = \varepsilon_r \varepsilon_0 \frac{A}{h}, \tag{3.80}$$

then, from (3.78–3.80) it follows that

$$\Delta V = P_Q \frac{A}{C_e} \Delta T = P_Q \frac{\varepsilon_r \varepsilon_0}{h} \Delta T \tag{3.81}$$

It is seen that the peak output voltage is proportional to the sensor's temperature rise and pyroelectric charge coefficient and inversely proportional to its thickness.

When the pyroelectric sensor is subjected to a thermal gradient its polarization (electric charge developed across the crystal) varies with the temperature of the crystal. A typical polarization-temperature curve is shown in Fig. 3.28. The voltage pyroelectric coefficient, P_v, is a slope of the polarization curve. It increases dramatically near the Curie temperature at which the polarization disappears and the material permanently loses its pyroelectric properties. The curves imply that the sensor's sensitivity increases with temperature at the expense of nonlinearity.

Piezo- and pyroelectric materials such as lithium tantalate and polarized ceramics are typical materials to produce the pyroelectric sensors. During recent years, a deposition of pyroelectric thin films have been intensively researched. Especially promising is use of lead titanate ($PbTiO_3$), which is a ferroelectric ceramic having both a high pyroelectric coefficient and a high Curie temperature of about 490°C. This material can be easily deposited on silicon substrates by the so called sol-gel spin casting deposition method [23].

Figure 3.29 shows the timing diagrams for a pyroelectric sensor when it is exposed to a step function of heat. It is seen that the electric charge reaches its peak value almost instantaneously, and then decays with a *thermal time constant*, τ_T. The physical meaning is this: a thermally induced polarization occurs initially in the most outer layer of the crystalline material (just few atomic layers), whose

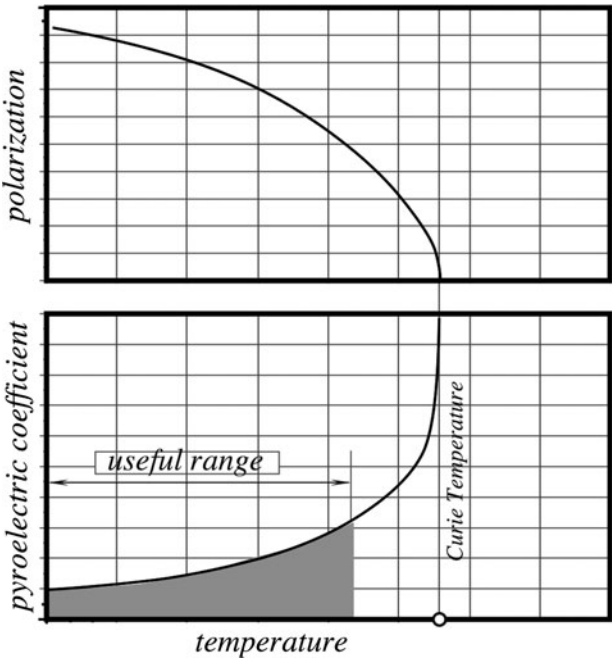

Fig. 3.28 Polarization of a pyroelectric crystal. The sensor must be stored and operated below the Curie temperature

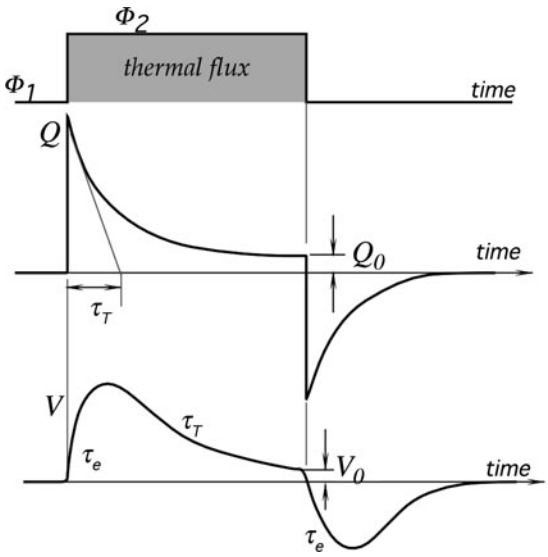

Fig. 3.29 Response of a pyroelectric sensor to a thermal step function The magnitudes of charge Q_0 and voltage V_0 are exaggerated for clarity

temperature nearly instantaneously raises to its maximum level. This creates the highest thermal gradient across the material thickness, leading to the maximum polarization. Then, heat propagates through the material, being absorbed by its mass in proportion to a thermal capacitance, C_T, and some of the heat is lost to the surroundings through a thermal resistance, R_T. This diminishes the initial gradient that generates the electric charge. The thermal time constant is a product of the sensors' thermal capacity and thermal resistance

$$\tau_T = C_T R_T = cAhR_T \tag{3.82}$$

where c is the specific heat of the pyroelectric element. The thermal resistance RT is function of all thermal losses to the surroundings through convection, conduction, and thermal radiation. For the low-frequency applications, it is desirable to use sensors with τ_T as large as practical, while for the high-speed applications (for instance, to measure laser pulses), a thermal time constant should be dramatically reduced. For that purpose, the pyroelectric material may be laminated with a heat sink: a piece of aluminum or copper.

When a pyroelectric sensor is exposed to a heat source, we consider a thermal capacity of the source being very large (an infinite heat source), and the thermal capacity of the sensor small. Therefore, the surface temperature T_b of a target can be considered constant during the measurement, while temperature of the sensor Ts is a function of time. That time function is dependent on the sensing element: its density, specific heat and thickness as per (3.82). If the input thermal flux has shape of a step function of time and the sensor is freely mounted in air, the output current can be approximated by an exponential function, so that

$$i = i_0 \mathrm{e}^{-t/\tau} \tag{3.83}$$

where i_0 is the peak current.

In Fig. 3.29, as long as heat source is present, the charge Q and voltage V do not completely return to zero, no matter how much time has elapsed. Thermal energy enters the pyroelectric material from side b (Fig. 3.26), resulting in a material temperature increase. This causes the sensor's response, which decays with a thermal time constant τ_T. However, since the other side a of the sensor faces a cooler environment, part of the thermal energy leaves the sensor and is lost to its surroundings. Because the sides a and b face objects of different temperatures (one is a temperature of a source and the other is a temperature of the environment), a continuous heat flow exists through the pyroelectric material. Electric current generated by the pyroelectric sensor has the same shape as the thermal current through its material. An accurate measurement can demonstrate that as long as the heat continues to flow, the pyroelectric sensor will generate a constant voltage v_0 whose magnitude is proportional to the heat flow, thus making the device a heat flow sensor.

3.8 Hall Effect

This physical effect was discovered in 1879 in Johns Hopkins University by E. H. Hall. Initially, the effect had a limited, however, a very valuable application as a tool for studying electrical conduction in metals, semiconductors, and other conductive materials. Nowadays, the Hall sensors are used to detect magnetic fields, position, and displacement of objects [24, 25].

The effect is based on the interaction between moving electric carriers and an external magnetic field. In metals, these carriers are electrons. When an electron moves through a magnetic field, upon it acts a sideways force

$$\mathbf{F} = q\mathbf{v}\mathbf{B}, \tag{3.84}$$

where q=1.6 × 10^{-19} C is an electronic charge, $\boldsymbol{v}$ is the speed of an electron, and **B** is the magnetic field. Vector notations (bold face) are an indication that the force direction and its magnitude depend on the spatial relationship between the magnetic field and the direction of the electron movement. The unit of **B** is 1 tesla=1 newton/(ampere·meter) = 10^4 gauss.

Magnetic field deflects movement of electric charges. Let us assume that the electrons move inside a flat conductive strip that is placed in magnetic field **B** (Fig. 3.30). The strip has two additional contacts at its left and right sides which are

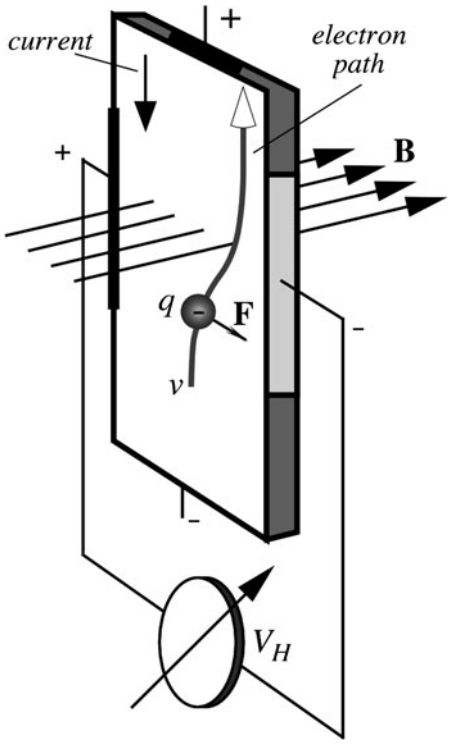

Fig. 3.30 Hall effect sensor. Magnetic field deflects movement of electric charges

connected to a voltmeter. Two other contacts are placed at the upper and lower ends of the strip. These are connected to a source of electric current. Due to the magnetic field, the deflecting force shifts moving electrons toward the right side of the strip which becomes more negative than the left side. That is, the magnetic field and the electric current produce the so called *transverse Hall potential difference* V_H. The sign and amplitude of this potential depends on both magnitude and directions of magnetic field and electric current. At a fixed temperature it is given by

$$V_H = hiB\sin\alpha, \tag{3.85}$$

where α is the angle between the magnetic field vector and the Hall plate (Fig. 3.31), and h is the coefficient of overall sensitivity whose value depends on the plate material, its geometry (active area), and its temperature.

The overall sensitivity depends on the *Hall coefficient*, which can be defined as the transverse electric potential gradient per unit magnetic field intensity per unit current density. According to the free electron theory of metals, the Hall coefficient should be given by

$$H = \frac{1}{Ncq}, \tag{3.86}$$

where N is the number of free electrons per unit volume and c is the speed of light. Depending on the material crystalline structure, charges may be either electrons (negative) or holes (positive). As a result, the Hall effect may be either negative or positive.

A linear Hall effect sensor is usually packaged in a four-terminal housing. Terminals for applying the control current are called the control terminals and a resistance between them is called the control resistance R_i. Terminals where the

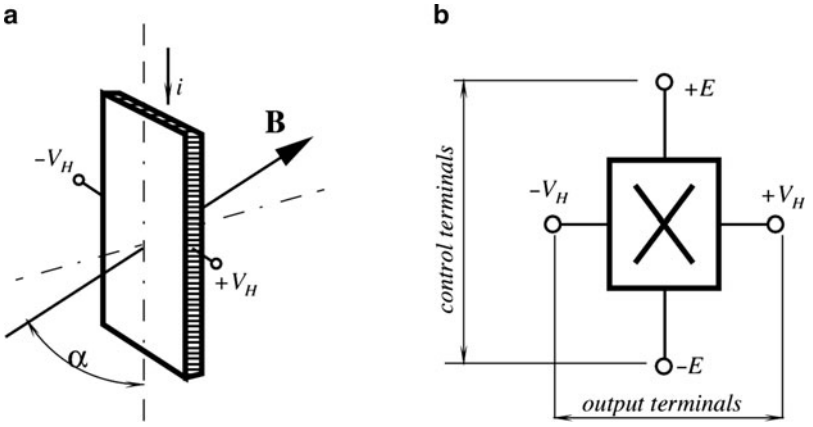

Fig. 3.31 Output signal of a Hall sensor depends on the angle between the magnetic field vector and the plate (**a**); four terminals of a Hall sensor (**b**)

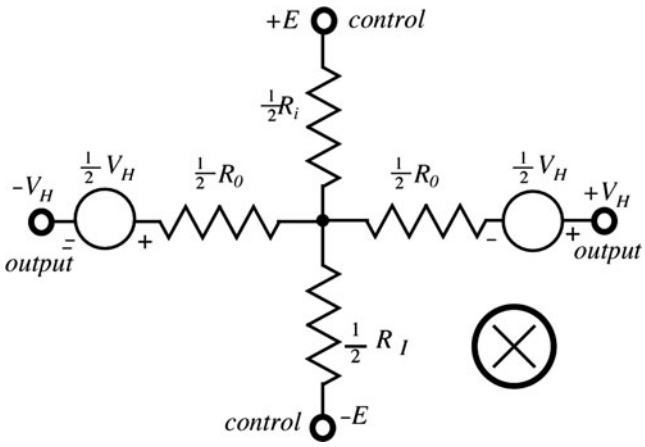

Fig. 3.32 Equivalent circuit of a Hall sensor

output voltage is observed are called the differential output terminals and a resistance between them is called the differential output resistance, R_o. The sensor's equivalent circuit (Fig. 3.32) may be represented by cross-connected resistors and two voltage sources connected in series with the output terminals. The cross ⊗ in Figs. 3.31b and 3.32 indicates the direction of the magnetic field from the viewer to the symbol plane.

The sensor is specified by its resistances, R_i and R_o, across both pairs of terminals, the offset voltage at no magnetic field applied, the sensitivity and the temperature coefficient of sensitivity. Many Hall effect sensors are fabricated from silicon and fall into two general categories: the basic sensors and the integrated sensors. Other materials used for the element fabrication include InSb, InAs, Ge, and GaAs. In the silicon element, an interface electronic circuit can be incorporated into the same wafer. This integration is especially important since the Hall effect voltage is quite small. As an example, see characteristics in Table 3.2 for a linear basic silicon sensor UGN-3605K manufactured by Sprague®.

A built-in electronic interface circuit may contain a threshold device, thus making an integrated sensor a two-state device. That is, its output is "zero" when the magnetic field is below the threshold, and it is "one" when the magnetic field is strong enough to cross the threshold.

Because of a piezoresistivity of silicon, all Hall effect sensors are susceptible to mechanical stress effects. Caution should be exercised to minimize the application of stress to the leads or the housing. The sensor is also sensitive to temperature variations because temperature influences resistance of the element. If the element is fed by a voltage source, temperature will change the control resistance, and subsequently the control current. Hence, it is preferable to connect the control terminals to a current source rather than to a voltage source.

One way to fabricate the Hall sensor is to use a silicon p-substrate with ion-implanted n-wells (Fig. 3.33a). Electrical contacts provide connections to the power

Table 3.2 Typical characteristics of a linear Hall Effect sensor (*Source*: [26])

Control current	3 mA
Control resistance, R_i	2.2 kΩ
Control resistance vs. temperature	+0.8%/°C
Differential output resistance, R_o	4.4 kΩ
Output offset voltage	5.0 mV (at B=0 Gauss)
Sensitivity	60 μV/Gauss
Sensitivity vs. temperature	+0.1%/°C
Overall sensitivity	20 V/ΩkG
Maximum magnetic flux density, B	Unlimited

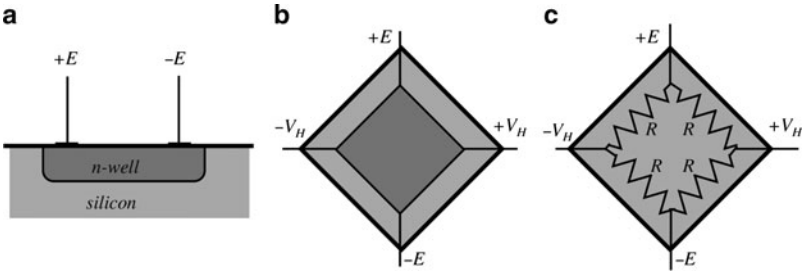

Fig. 3.33 Silicon Hall effect sensor with n-well (**a** and **b**) and equivalent resistive bridge circuit (C)

supply terminals and form the sensor outputs. A Hall element is a simple square where a well with four electrodes attached to the diagonals (Fig. 3.33b). A helpful way of looking at the Hall sensor is to picture it as a resistive bridge depicted in Fig. 3.33c. This representation makes its practical applications more conventional because the bridge circuits are the most popular networks with well-established methods of design (Sect. 5.10).

3.9 Thermoelectric Effects

3.9.1 Seebeck Effect

In 1821, Thomas Johann Seebeck (1770–1831), an Estonian born and Berlin and Göttingen educated physician, accidentally joined semicircular pieces of bismuth and copper while studying thermal effects on galvanic arrangements [27]. A nearby compass indicated a magnetic disturbance (Fig. 3.34a). Seebeck experimented repeatedly with different metal combinations at various temperatures, noting related magnetic field strengths. Curiously, he did not believe that an electric current was flowing, and preferred to describe that effect as "thermomagnetism" [28].

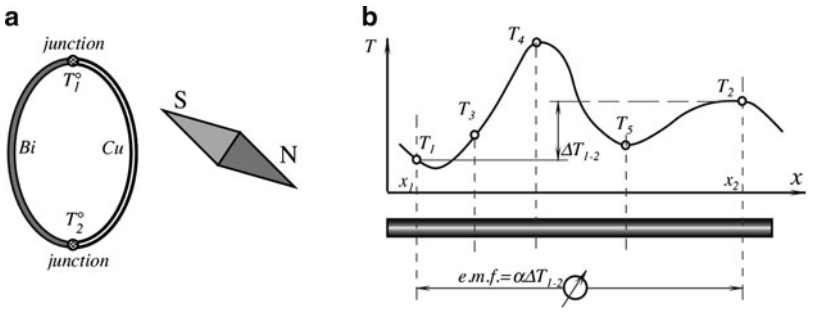

Fig. 3.34 Seebeck experiment (**a**), varying temperature along a conductor is a source of a thermoelectric e.m.f. (**b**)

If we take a conductor and place one end of it into a cold place and the other end into a warm place, energy will flow from the warm to cold part. The energy takes the form of heat. The intensity of the heat flow is proportional to the thermal conductivity of the conductor. Besides, the thermal gradient sets an electric field inside the conductor (this directly relates to Thompson effect).[12] The field results in incremental voltage

$$dV_a = \alpha_a \frac{dT}{dx} dx, \tag{3.87}$$

where dT is the temperature gradient across small length dx and α_a is the *absolute* Seebeck coefficient of the material [29]. If the material is homogeneous, α_a is not a function of length and (3.87) reduces to

$$dV_a = \alpha_a dT. \tag{3.88}$$

Equation (3.88) is a principle mathematical expression of a thermoelectric effect. Figure 3.34b shows a conductor having nonuniform temperature T along its length x. A temperature gradient between any arbitrary points defines an electromotive force (e.m.f.) between these points. Other possible temperatures between the selected points (temperatures T_3, T_4, and T_5, for example) make no effect whatsoever on the value of e.m.f. between points 1 and 2.

To measure e.m.f., we would like to connect a voltmeter to the conductor as shown in Fig. 3.34b and this is not as simple as may first look. To measure thermally induced e.m.f. we would need to attach the voltmeter probes. However, the probes are also made of conductors that may be different from the conductor we study. As

[12] A Thompson effect was discovered by William Thompson around 1850. It consists of absorption or liberation of heat by passing current through a homogeneous conductor which has a temperature gradient across its length. Unlike in the Joule effect, the heat is linearly proportional to current. Heat is absorbed when the current and heat flow in opposite directions, and heat is produced when they flow in the same direction.

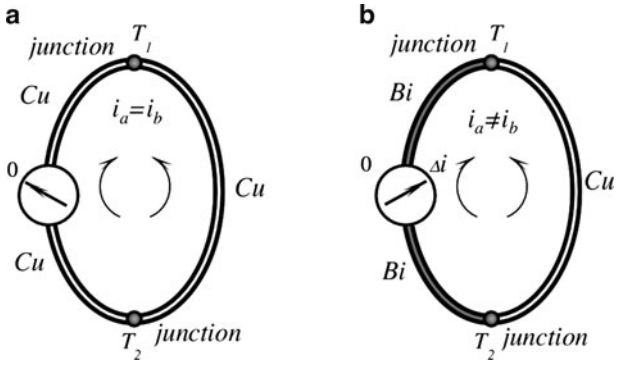

Fig. 3.35 Thermoelectric loop
Joints of identical metals produce zero net current at any temperature difference (**a**); joints of dissimilar metals produce net current Δi (**b**)

a result, the probe contacts will introduce their own e.m.f. and disturb out experiment. Let us consider a simple measurement electric circuit where a current loop is formed. We cut the left side of the conductor (Cu) and insert the current meter into the cut in series with the wire (Fig. 3.35a). If the entire loop is made of a uniform material, say cooper, then no current will be observed, even if the temperature along the conductor is not uniform. Electric fields in the left and right arms of the loop produce equal currents $i_a=i_b$ which cancel one another, resulting in zero net current. A thermally-induced e.m.f. exists in every thermally non-homogeneous conductor, but it ca nnot be directly measured.

In order to observe thermoelectricity, it is in fact necessary to have a circuit composed of two different materials,[13] and we can then measure the net difference between their thermoelectric properties. Figure 3.35b shows a loop of two dissimilar metals which produces net current $\Delta i=i_a-i_b$. The actual current depends on many factors, including the shape and size of the conductors. If, on the other hand, instead of current we measure the net voltage across the broken conductor, the potential will depend *only* on the materials and the temperature difference. It does not depend on any other factors. A thermally induced potential difference is called the Seebeck potential. Note that the only way to eliminate influence of the voltmeter terminals is to connect it into a cut as shown in Fig. 3.35a, that is, both terminals of the voltmeter must be connected to the same type of a conductor.

What happens when two conductors are joined together? Free electrons in metal may behave as an ideal gas. Kinetic energy of electrons is a function of the material temperature. However, in different materials, energies and densities of free electrons are not the same. When two dissimilar materials at the same temperature are brought into a contact, free electrons diffuse through the junction [29].

[13]2 Or perhaps the same material in two different states, for example, one under strain, the other is not.

The electric potential of the material accepting electrons becomes more negative at the interface, while the material emitting electrons becomes more positive. Different electronic concentrations across the junction set up an electric field, which balances the diffusion process and the equilibrium is established. If the loop is formed and both junctions are at the same temperature, the electric fields at both junctions cancel each other, which is not the case when the junctions are at different temperatures.

A subsequent investigation [30] has shown the Seebeck effect to be fundamentally electrical in nature. It can be stated that the thermoelectric properties of a conductor are in general just as much bulk properties as are the electrical and thermal conductivities. Coefficient α_a is a unique property of a material. When a combination of two dissimilar materials (A and B) is used, the Seebeck potential is determined from a differential Seebeck coefficient:

$$\alpha_{AB} = \alpha_A - \alpha_B \tag{3.89}$$

and the net voltage of the junction is

$$dV_{AB} = \alpha_{AB} dT. \tag{3.90}$$

The above equation can be used to determine a differential coefficient

$$\alpha_{AB} = \frac{dV_{AB}}{dT} \tag{3.91}$$

Note that the Seebeck coefficient is not really a constant. It is temperature dependent and thus the Seebeck potential will be different at different temperatures. For example, voltage as function of a temperature gradient for a T-type thermocouple with a high degree of accuracy can be approximated by a second order equation

$$V_{AB} = a_0 + a_1 T + a_2 T^2 = -0.0543 + 4.094 \times 10^{-2} T + 2.874 \times 10^{-5} T^2 \tag{3.92}$$

then a differential Seebeck coefficient for the T-type thermocouple is

$$\alpha_T = \frac{dV_{AB}}{dT} = \alpha_1 + 2\alpha_2 = 4.094 \times 10^{-2} + 5.7481 \times 10^{-5} T \tag{3.93}$$

It is seen that the coefficient is a linear function of temperature. Sometimes this coefficient α_{AB} is called the sensitivity of a thermocouple junction. A junction, which is kept at a cooler temperature, traditionally is called a cold junction and the warmer is a hot junction. The Seebeck coefficient does not depend on the nature of the junction: metals may be pressed together, welded, fused, twisted, etc. What counts is the temperature of the junction and the actual metals. The Seebeck effect is a direct conversion of thermal energy into electric energy.

Table A.11 in Appendix gives values of thermoelectric coefficients and volume resistivities for some thermoelectric materials. It is seen that to achieve the best sensitivity, the junction materials shall be selected with the opposite signs for α and those coefficients should be as large as practical.

In 1826, A. C. Becquerel suggested to use the Seebeck's discovery for temperature measurements. Nevertheless, the first practical thermocouple was constructed by Henry LeChatelier almost 60 years later [31]. He had found that the junction of platinum and platinum-rhodium alloy wires produce "the most useful voltage." Thermoelectric properties of many combinations have been well documented and for many years used for measuring temperature. Table A.10 (Appendix) gives sensitivities of some practical thermocouples (at 25°C) and Fig. 3.36 shows the Seebeck voltages for the standard types of thermocouples over a broad temperature range. It should be emphasized once again that thermoelectric sensitivity is not constant over the temperature range and it is customary to reference thermocouples at 0°C. Besides the thermocouples, the Seebeck effect also is employed in thermopiles, which are, in essence, multiple serially connected thermocouples. Nowadays, thermopiles are most extensively used for the detection of thermal radiation (Sect. 14.7.2). The original

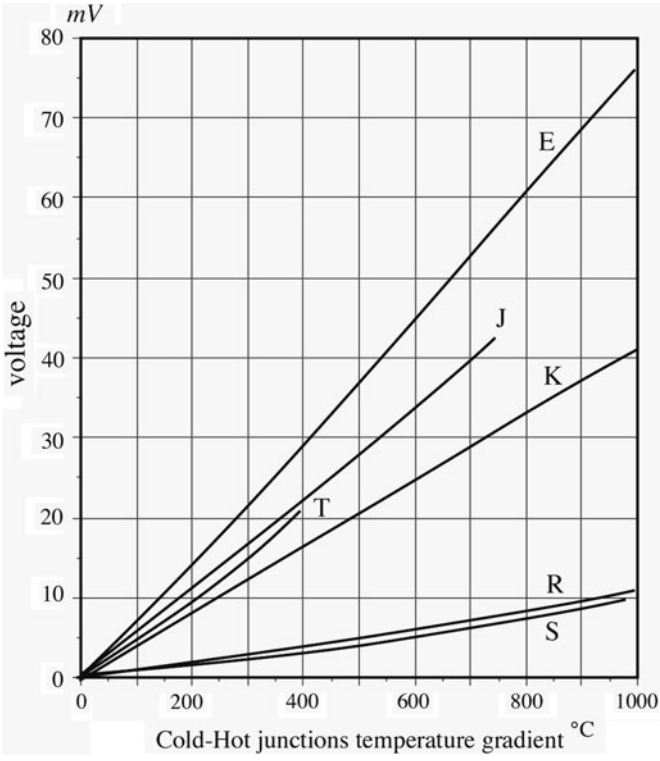

Fig. 3.36 Output voltage from standard thermocouples as functions of a cold-hot temperature gradient

thermopile was made of wires and intended for increasing the output voltage. It was invented by James Joule (1818–1889) [32].

Nowadays, the Seebeck effect is used in fabrication of the integral sensors where pairs of materials are deposited on the surface of semiconductor wafers. An example is a thermopile, which is a sensor for detection of thermal radiation. Quite sensitive thermoelectric sensors can be fabricated of silicon, since silicon possesses a strong Seebeck coefficient. The Seebeck effect results from the temperature dependence of the Fermi energy E_F, and the total Seebeck coefficient for n-type silicon may be approximated as a function of electrical resistivity for the range of interest (for use in sensors at room temperature):

$$\alpha_a = \frac{mk}{q} \ln \frac{\rho}{\rho_0}, \tag{3.94}$$

where $\rho_0 \approx 5 \times 10^{-6}$ Ωm and $m \approx 2.5$ are constants, k is the Boltzmann constant, and q is the electronic charge. The doping concentrations used in practice lead to Seebeck coefficients on the order of 0.3–0.6 mV/K. The absolute Seebeck coefficients of a few selected metals and some typical values of silicon are shown in Table A.11. It can be seen that the Seebeck coefficients for metals are much smaller than for silicon and that the influence of aluminum terminals on chips is negligible compared to the Seebeck coefficient for silicon. We conclude this discussion of the thermoelectric effect by a remark that the effect allows fabrication a relative temperature sensor but not an absolute sensor. In other words, a thermocouple or thermopile sensor will measure only a temperature gradient. To measure an absolute temperature, a cold or hot junction temperature must be either known or measured by another sensor, the reference absolute sensor, such as a thermistor, for example.

3.9.2 Peltier Effect

In the early nineteenth century, a French watchmaker-turned-physicist, Jean Charles Athanase Peltier (1785–1845) discovered that if electric current passes from one substance to another (Fig. 3.37) and then heat may be given or absorbed at

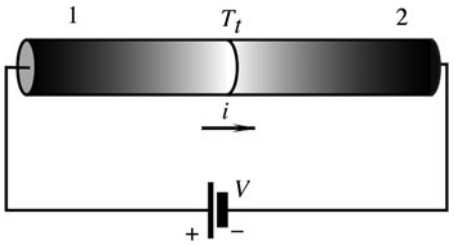

Fig. 3.37 Peltier effect

the junction [33]. Heat absorption or production is a function of the current direction

$$\mathrm{d}Q_{\mathrm{P}} = \pm pi\,\mathrm{d}t, \tag{3.95}$$

where i is the current and t is time. The coefficient p has a dimension of voltage and represents thermoelectric properties of the material. It should be noted that heat does not depend on temperature at the other junction.

The Peltier effect concerns the reversible absorption of heat, which usually takes place when an electric current crosses a junction between two dissimilar metals. The effect takes place whether the current is introduced externally or is induced by the thermocouple junction itself (due to Seebeck effect).

The Peltier effect is used for two purposes: it can produce heat or "produce" cold (remove heat), depending on the direction of electric current through the junction. This makes it quite useful for the devices where precision thermal control is required. Apparently, the Peltier effect is of the same nature as the Seebeck effect. It should be well understood that the Peltier heat is different from that of the Joule. The Peltier heat depends linearly on the magnitude of the current flow as contrasted to Joule heat.[14] The magnitude and direction of Peltier heat do not depend in any way on the actual nature of the contact. It is purely a function of two different bulk materials, which have been brought together to form the junction and each material makes its own contribution depending on its thermoelectric properties. The Peltier effect is a basis for operation of thermoelectric coolers, which are used for cooling of the photon detectors operating in the far infrared spectral range (Sect. 14.5) and chilled mirror hygrometers (Sect. 13.5).

In summary, thermoelectric currents may exist whenever the junctions of a circuit formed of at least two dissimilar metals are exposed to different temperatures. This temperature difference is always accompanied by irreversible Fourier heat conduction, while the passage of electric currents is always accompanied by irreversible Joule heating effect. At the same time, the passage of electric current always is accompanied by reversible Peltier heating or cooling effects at the junctions of the dissimilar metals, while the combined temperature difference and passage of electric current always is accompanied by reversible Thomson heating or cooling effects along the conductors. The two reversible heating-cooling effects are manifestations of four distinct e.m.f.'s which make up the net Seebeck e.m.f.

$$E_{\mathrm{s}} = p_{AB_{T2}} - p_{AB_{T1}} + \int_{T_1}^{T_2} \sigma_A dT - \int_{T_1}^{T_2} \sigma_B dT = \int_{T_1}^{T_2} \alpha_{AB} dT, \tag{3.96}$$

[14] Joule heat is produced when electric current passes in any direction through a conductor having finite resistance. Released thermal power of Joule heat is proportional to squared current: $P = i^2/R$, where R is resistance of a conductor.

where σ is a quantity called the Thomson coefficient, which Thomson referred to as the specific heat of electricity because of an apparent analogy between σ and the usual specific heat, c of thermodynamics. The quantity of σ represents the rate at which heat is absorbed, or liberated, per unit temperature difference per unit mass [34, 35].

3.10 Sound Waves

An alternate physical compression and expansion of medium (solids, liquids, and gases) with certain frequencies are called sound waves. The medium contents oscillate in the direction of wave propagation, hence these waves are called longitudinal mechanical waves. The name *sound* is associated with the hearing range of a human ear, which is approximately from 20 to 20,000 Hz. Longitudinal mechanical waves below 20 Hz are called *infrasound* and above 20,000 Hz (20 kHz) *ultrasound*. If the classification were made by other animals, like dogs, the range of sound waves surely would be wider.

Detection of infrasound is of interest with respect to analysis of building structures, earthquake prediction, and other geometrically large sources. When infrasound is of a relatively strong magnitude, it can be if not heard, at least felt by humans, producing quite irritating psychological effects (panic, fear, etc.).[15] Audible waves are produced by vibrating strings (string music instruments), vibrating air columns (wind music instruments), and vibrating plates (some percussion instruments, vocal cords, loudspeakers). Whenever sound is produced, air is alternatively compressed and rarefied. These disturbances propagate outwardly. A spectrum of waves may be quite different - from a simple monochromatic sounds from a metronome or an organ pipe, to a reach multiharmonic violin sound. Acoustic noise may have a very broad spectrum. It may be of a uniform distribution of density or it may be "colored" with predominant harmonics at some of its portions.

When a medium is compressed, its volume changes from V to $V - \Delta V$. The ratio of change in pressure, Δp, to relative change in volume is called the bulk modulus of elasticity of medium:

$$B = -\frac{\Delta p}{\Delta V/V} = \rho_0 v^2, \tag{3.97}$$

where ρ_0 is the density outside the compression zone and v is the speed of sound in the medium. Then speed of sound can be defined as

[15]There is an anecdote about the American physicist R. W. Wood (1868–1955). A theatrical director from New York asked Wood to invent a mysterious sound effect for a play about travel through time. Wood built a huge organ pipe (sort of a whistle) for the infrasonic frequency of about 8 Hz. When during a dress-rehearsal Wood activated the pipe, the entire building and everything in it started vibrating. The terrified audience ran out to the street, feeling uncontrollable fear and panic. Needless to say, the pipe was never used during performances.

$$\mathrm{v} = \sqrt{\frac{B}{\rho_0}}, \tag{3.98}$$

Hence, the speed of sound depends on the elastic (B) and inertia (ρ_0) properties of the medium. Since both variables are functions of temperature, the speed of sound also depends on temperature. This feature forms a basis for operation of the acoustic thermometers (Sect. 16.5). For solids, longitudinal velocity can be defined through its Young modulus E and Poisson ratio v:

$$\mathrm{v} = \sqrt{\frac{E(1-v)}{\rho_0(1+v)(1-2v)}} \tag{3.99}$$

Table A.15 (Appendix) provides speeds of longitudinal waves in some media.

If we consider the propagation of a sound wave in an organ tube, each small volume element of air oscillates about its equilibrium position. For a pure harmonic tone, the displacement of a particle from the equilibrium position may be represented by

$$y = y_m \cos\frac{2}{\lambda}(x - \mathrm{v}t), \tag{3.100}$$

where x is the equilibrium position of a particle and y is a displacement from the equilibrium position ym is the amplitude, and λ is the wavelength. In practice, it is more convenient to deal with pressure variations in sound waves rather than with displacements of the particles. It can be shown that the pressure exerted by the sound wave is

$$p = k\rho_0 v^2 y_\mathrm{m} \sin(kx - \omega t), \tag{3.101}$$

where $k = 2\pi/\lambda$ is a wave number, ω is angular frequency, and the front terms represent an amplitude, p_m, of the sound pressure. Therefore, a sound wave may be considered as a pressure wave. It should be noted that sin and cos in (3.100) and (3.101) indicate that the displacement wave is 90° out of phase with the pressure wave.

Pressure at any given point in media is not constant and changes continuously, and the difference between the instantaneous and the average pressure is called an acoustic pressure P. During the wave propagation, vibrating particles oscillate near a stationary position with the instantaneous velocity ξ. The ratio of the acoustic pressure and the instantaneous velocity (do not confuse it with a wave velocity) is called an acoustic impedance:

$$\mathbf{Z} = \frac{\mathbf{P}}{\xi}, \tag{3.102}$$

which is a complex quantity, which is characterized by an amplitude and a phase. For an idealized media (no loss) the **Z** is real and is related to the wave velocity as

$$Z = \rho_0 v. \tag{3.103}$$

We can define *intensity I* of a sound wave as the power transferred per unit area. Also, it can be expressed through the acoustic impedance

$$I = P\xi = \frac{P^2}{Z}. \tag{3.104}$$

It is common, however, to specify sound not by intensity but rather by a related parameter β, called the sound level and defined with respect to a reference intensity $I_0 = 10^{-12}$ W/m^2:

$$\beta = 10\log_{10}\left(\frac{I}{I_0}\right) \tag{3.105}$$

The magnitude of I_0 was chosen because it is the lowest ability of a human ear. The unit of β is a decibel (dB), named after Alexander Graham Bell. If $I = I_0$, $\beta = 0$.

Pressure levels also may be expressed in decibels as

$$\Pi = 20\log_{10}\left(\frac{p}{p_0}\right), \tag{3.106}$$

where $p_0 = 2 \times 10^{-5}$ N/m^2.

Table 3.3 Sound levels (β) referenced to I_0 at 1,000 Hz

Sound source	dB
Rocket engine at 50 m	200
Supersonic boom	160
Hydraulic press at 1 m	130
Threshold of pain	120
10W Hi-Fi speaker at 3 m	110
Unmuffled motorcycle	110
Rock-n-roll band	100
Subway train at 5 m	100
Pneumatic drill at 3 m	90
Niagara Falls	85
Heavy traffic	80
Automobiles at 5 m	75
Dishwashers	70
Conversation at 1 m	60
Accounting office	50
City street (no traffic)	30
Whisper at 1 m	20
Rustle of leaves	10
Threshold of hearing	0

Examples of some sound levels are given in Table 3.3. Since the response of a human ear is not the same at all frequencies, sound levels are usually referenced to I_0 at 1 kHz where the ear is most sensitive.

3.11 Temperature and Thermal Properties of Materials

Our bodies have a sense of temperature, which by no means is an accurate method to measure outside heat. Human senses are not only nonlinear, but relative with respect to our previous experience. Nevertheless, we can easily tell the difference between warmer and cooler objects. Then, what is going on with these objects that they produce different perceptions?

Every single particle in this universe exists in perpetual motion. The temperature of a volume of a material, in the simplest way, can be described as measure of an average kinetic energy of vibrating particles. The stronger the movement the higher the temperature. Of course, molecules and atoms in a given volume of material do not move with equal intensities. That is, microscopically, they all are at different temperatures. The average kinetic energy of a large number of moving particles determines macroscopic temperature of an object. These processes are studied by a thermodynamics and statistical mechanics. Here, however, we are concerned with methods and devices that are capable of measuring the macroscopic average kinetic energy of vibrating particles, which is the other way to say the temperature of the object. Since temperature is related to the movement of molecules, it is closely associated with pressure, which is defined as the force applied by moving molecules per unit area.

When atoms and molecules in a material move, they interact with other molecules that happen to be brought in contact with them. A jiggling atom agitates a neighboring atom and transfers to it portion of its kinetic energy, so the neighbor starts vibrating more intensely. This agitation propagates through the material, elevating the temperature. Since an atom contains moving electrons swirling around its nucleus like a cloud of electric current, thermal agitation causes that current to produce electromagnetic wave. Hence, every vibrating atom acts as a microscopic radio-transmitter, which emanates electromagnetic radiation to the surrounding space. These two types of activities form a basis for heat transfer from warmer to cooler objects: conduction and radiation. The stronger the atomic movement the hotter the temperature and the stronger the electromagnetic radiation. A special device (we call it a thermometer), which either contacts the object or receives its electromagnetic radiation, then produces a physical response, or signal. That signal becomes a measure of the object's temperature.

The word thermometer first appeared in literature in 1624 in a book by J. Leurechon, entitled La Récréation Mathématique [29]. The author described a glass water-filled thermometer whose scale was divided by 8°. The first

pressure-independent thermometer was built in 1654 by Ferdinand II, Grand Duke of Tuscany[16] in form of an alcohol-filled hermetically sealed tube.

Thermal energy is what we call heat. Heat is measured in calories.[17] One calorie (cal) is equal to amount of heat, which is required to warm up by 1°C 1 g of water at normal atmospheric pressure. In the United States, a British unit of heat is generally used, which is 1 Btu (British thermal unit): 1 Btu=252.02 cal.

3.11.1 Temperature Scales

There are several scales to measure temperature. To make a linear scale (for convenience, all thermometers have linear scales), at least two reference points are required. Usually one of these points is called a zero point. A first zero for a scale was established in 1664 by Robert Hooke at a point of freezing distilled water. In 1694 Carlo Renaldi of Padua suggested to take a melting point of ice (zero point) and a boiling point of water (second point) to define a span of his thermometer. He divided the span by 12 equal parts. Unfortunately, his suggestion had been forgotten for almost 50 years. In 1701, Newton also suggested for the zero point to use the temperature of melting ice and for the second point he chose the armpit temperature of a "healthy Englishman," he labeled that point "12." At Newton's scale, water was boiling at point No. 34. Daniel Gabriel Fahrenheit, a Dutch instrument maker, in 1706 selected zero for his thermometer at the coldest temperature he could produce by a mixing water, ice, and sal-ammoniac or household salt. For the sake of convenience, he established the other point at 96°, which was "found in the blood of a healthy man."[18] On his scale, the melting point of water was at 32° and boiling at 212°. In 1742, Andreas Celsius, professor of astronomy at the University of Uppsala (Sweden), proposed a scale with zero as the melting point of ice and 100° at boiling point of water. He divided the span by 100 equal parts – degrees.

Nowadays, in science and engineering, Celsius and Kelvin scales are generally employed. The Kelvin scale is arbitrarily based on the so-called triple point of water. There is a fixed temperature at a unique pressure of 4.58 mmHg where water vapor, liquid, and ice can coexist. This unique temperature is 273.16 K (degrees kelvin), which approximately coincides with 0°C. The Kelvin scale is linear with zero intercept (0 K) at a lowest temperature where kinetic energy of all moving

[16]More precisely, not "*by him*" but rather *"for him"*.

[17]A *calorie* that measures energy in food is actually equal to 1,000 physical calories, which is called a *kilocalorie*.

[18]After all, Fahrenheit was a toolmaker and for him 96 was a convenient number because to engrave the graduation marks, he could easily do so by dividing a distance between the marks by two: 96, 48, 24, etc. With respect to nationality of the blood, he did not care if it was blood of an Englishman or not. Now, it is known that blood temperature of a healthy person is not really constant. It varies between approximately 97°F and 99.5°F (36°C and 37.5°C) but during his times, Fahrenheit could not find a better thermostat than a human body.

particles is equal to zero. This point cannot be exactly attained in practice and is a strictly theoretical limit. It is called the absolute zero. Kelvin and Celsius scales have the same slopes,[19] i.e., 1°C=1 K and 0 K = −273.15°C. So the Kelvin scale is a shifted Celsius scale:

$$^{\circ}\mathrm{C} = {}^{\circ}\mathrm{K} - 273.15^{\circ} \tag{3.107}$$

The boiling point of water is at 100°C = 373.15°K. In the past, the Celsius scale sometimes was called "centigrade scale." Now, this term is no longer in use.

A slope of the Fahrenheit scale is steeper, because 1°C = 1.8°F. The Celsius and Fahrenheit scales cross at −40°C and F. The conversion between the two scales is

$$^{\circ}\mathrm{F} = 32 + 1.8^{\circ}\mathrm{C} \tag{3.108}$$

which means that at 0°C, temperature on the Fahrenheit scale is +32°F.

3.11.2 Thermal Expansion

Essentially, all solids expand in volume with an increase in temperature. This is a result of vibrating atoms and molecules. When the temperature goes up, an average distance between the atoms increases, which leads to an expansion of a whole body.[20] The change in any linear dimension: length, width, or height is called a *linear expansion*. A length, l_2, at temperature, T_2, depends on length, l_1, at initial temperature T_1:

$$l_2 = l_1[1 + (T_2 - T_1) \tag{3.109}$$

where α, called the coefficient of linear expansion, has different values for different materials. It is defined as

$$\alpha = \frac{\Delta l}{l}\frac{1}{\Delta T} \tag{3.110}$$

where $\Delta T = T_2 - T_1$. Table A.16 in Appendix gives values of α for different materials.[21] Strictly speaking, α depends on the actual temperature. However, for most engineering purposes, small variations in α may be neglected. For the

[19]There is a difference of 0.01° between the Kelvin and Celsius scales, as Celsius' zero point is defined not at a triple point of water as for the Kelvin, but at temperature where ice and air-saturated water are at equilibrium at atmospheric pressure.

[20]This assumes that there is no phase change during warming up, like from solid to liquid.

[21]More precisely, thermal expansion can be modeled by higher order polynomials, however, for the majority of practical purposes, a linear approximation is usually sufficient.

so-called *isotropic* materials, α is the same for any direction. The fractional change in area of an object and its volume with a high degree of accuracy can be represented, respectively, by

$$\Delta A = 2\alpha A \Delta T, \tag{3.111}$$

$$\Delta V = 3\alpha V \Delta T, \tag{3.112}$$

Thermal expansion is a useful phenomenon that can be employed in many sensors where thermal energy is either measured or used as an excitation signal. Consider two laminated plates, X and Y, that are fused together (Fig. 3.38a). The plates have the same thickness, surface areas, and identical moduli of elasticity. Their coefficients of thermal expansion, α_1 and α_2, however, are different. The fused plates are anchored at the left-hand side to the reference wall. Now, if we apply heat to the structure, that is, if we increase its temperature from T_1 to T_2, plate X will expand more than plate Y (for $\alpha_1 > \alpha_2$). The joint between the plates will restrain plate X from a uniform expansion, while forcing plate Y to expand more, than its coefficient of expansion would require. This results in formation of the internal stress and the structure will warp downwardly. Contrary, if we cool the structure, it will warp upwardly. The radius of warping can be estimated from equation [36]:

$$r \approx \frac{2j}{3(\alpha_X - \alpha_Y)(T_2 - T_1)} \tag{3.113}$$

The warping results in deflection of the nonanchored portion of the laminated plates. The deflection is strongest at the end of the structure. This deflection can me measured as a representative of the temperature change with respect to the reference temperature (we may call it calibration temperature). At the calibration temperature the plate is flat, however, any convenient shape at a calibration temperature may be selected. A bimetal plate is a transducer of temperature into a displacement but not sensor since it does not produce electric output signal.

Most of such transducers are made of the bimetal plates using iron-nickel-chrome alloys. They are useful in a temperature range from $-75°C$ and up to $+600°C$. For relatively small temperature changes, radius of curvature is quite large (several meters) and thus the tip deflection is rather small. A bimaterial plate tip deflection can be computed from

$$\Delta = r\left(1 - \cos\frac{180L}{\pi r}\right), \tag{3.114}$$

where r is found from (3.113) and L is the length of the plate. For example, for a bimetal plate made of brass (α=20 $\times$ 10^{-6}) and chromium (α=6 $\times$ 10^{-6}) having L=50 mm and j=1 mm, for a 10°C gradient the deflection $\Delta \approx 0.26$ mm. This

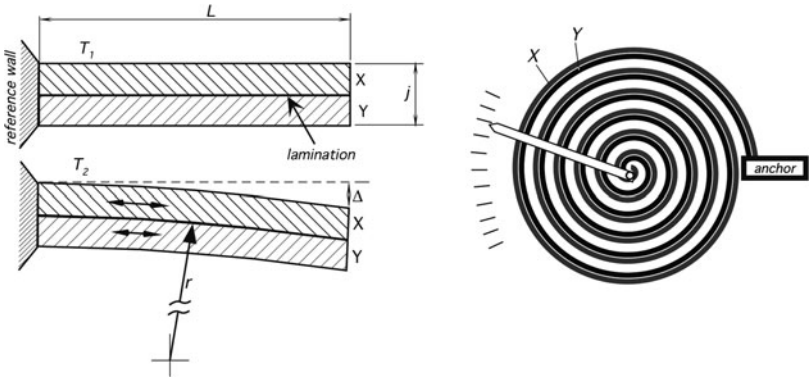

Fig. 3.38 Warping of a laminated plate where two materials have different coefficients of thermal expansion (**a**); bi-metal coil used as a temperature transducer (**b**)

deflection is not easy to observe with a naked eye, thus, in a practical thermometer a bimetal plate is usually preshaped in form of a coil (Fig. 3.38b). This allows for a dramatic increase in L and achieve a much larger Δ. In the same example, for L=200 mm, the deflection becomes 4.2 mm, which is a significant improvement. In modern sensors, the bimaterial structure is fabricated by employing a micromachining technology (MEMS).

3.11.3 Heat Capacity

When an object is warmed up, its temperature increases. By warming we mean transfer of a certain amount of heat (thermal energy) into the object. Heat is stored in the object in form of a kinetic energy of vibration atoms. Since different materials are composed of atoms having different atomic weights, which even may be locked into crystalline structures, the kinetic energy of the atomic vibration also will be different. The amount of heat which an object can store is analogous to the amount of water that a water tank can store. Naturally, it cannot store more than its volume, which is a measure of a tank's capacity. Similarly, every object may be characterized by a heat capacity, which depends on both the material properties of the object and its mass, m:

$$C = cm, \tag{3.115}$$

where c is a constant that characterizes thermal properties of material. It is called the *specific heat* and is defined as

$$c = \frac{Q}{m\Delta T} \tag{3.116}$$

The specific heat describes the material while a thermal capacity describes an object, which is made of that material. Strictly speaking, specific heat is not constant over an entire temperature range of a specific phase of the material. It may change dramatically when a phase of the material changes, say from solid to liquid. Microscopically, specific heat reflects structural changes in the material. For instance, the specific heat of water is almost constant between 0°C and 100°C (liquid phase). Almost, but not exactly- it is higher near freezing, and decreases slightly when the temperature goes to about 35°C and then slowly rises again from 38° to 100°. Remarkably, the specific heat of water is the lowest near 37°C, a biologically optimal temperature of the warm-blooded animals.[22]

Table A.17 gives the specific heat for various materials in cal/(g°C). Some other tables provide specific heat in SI units of energy which is joule/g °C. The relationship between cal/(g°C) and j/(g°C) is as follows:

$$1\frac{\mathrm{j}}{\mathrm{g^oC}} = 0.2388\frac{\mathrm{cal}}{\mathrm{g^oC}}. \tag{3.117}$$

It may be noted that generally the heavier the material the lower is its specific heat. A concept of thermal capacity is very important for development of temperature sensors. It follows from (3.116) that for the same temperature increase, a smaller thermal energy need be transferred from the object to a lighter sensor with smaller specific heat. Hence, a temperature sensor having smaller heat capacity will disturb the measured object to a lesser degree.

3.12 Heat Transfer

There are two fundamental properties of heat, which should be well recognized:
1. The heat is totally not specific, that is, once it is produced, it is impossible to say what origin it has, and
2. The heat can not be contained, which means that it flows spontaneously from warmer to the cooler part of the system and there is no method known to modern science to stop the heat flow entirely.

Thermal energy may be transferred from one object to another by three ways: conduction, convection, and radiation. While conduction and radiation relate to solids, liquids, and gases, convection can transfer heat by an intermediate fluid (liquid or gas). A measurement of any physical quantity always requires transfer of energy. Naturally, one of the objects involved in thermal exchange may be a thermal sensor. Its purpose would be to measure the amount of heat, which represents certain

[22]Likely, this is because of a better compatibility between the animal protein molecules and structures of the water crystals at that temperature.

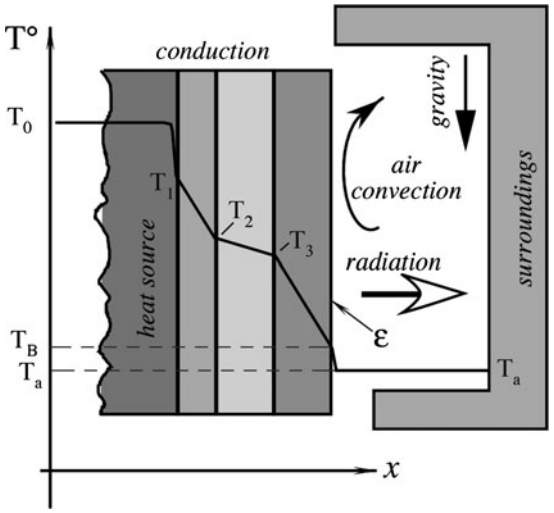

Fig. 3.39 Temperature profile in laminated materials

information about the object producing that heat. Such information may be the temperature of an object, chemical reaction, position of the object, etc.

Let us consider a sandwich-like multilayer object, where each layer is made of a different material. When heat moves through the layers, a temperature profile within each material depends on its thickness and thermal conductivity. Figure 3.39 shows three laminated layers where the first layer is attached to a heat source (a device having an "infinite" heat capacity and a high thermal conductivity). One of the best solid materials to act as an infinite heat source is a thermostatically controlled bulk copper. In liquids, an "infinite" heat capacity can be attributed to a stirred liquid, like water in a temperature-controlled bath. Temperature within the source is higher and constant, except of a thin boundary region at the joint of the laminated materials. Heat propagates from one material to another by conduction, gradually dropping toward ambient temperature. The temperature within each material drops with different rates depending on the thermal properties of the material. The last layer loses heat to air through the natural (gravitational) convection and to the surrounding objects through the infrared (thermal) radiation. Thus, Fig. 3.39 illustrates all three possible ways to transfer heat from one object to another: conduction, convection, and radiation.

3.12.1 Thermal Conduction

Heat conduction requires a physical contact between two bodies. Thermally agitated particles in a warmer body jiggle and transfer kinetic energy to a cooler body by agitating its particles. As a result, the warmer body loses heat while the cooler body gains heat. Heat transfer by conduction is analogous to water flow or to

electric current. For instance, heat passage through a rod is governed by a law, which is similar to Ohm's law. A heat flow rate (thermal "current") is proportional to a thermal gradient (thermal "voltage") across the material (dT/dx) and a cross-sectional area A:

$$H = \frac{dQ}{dt} = -kA\frac{\Delta T}{dx}, \tag{3.118}$$

where k is called thermal conductivity. The minus sign indicates that heat flows in the direction of temperature decrease (a negative derivative is required to cancel the minus sign). A good thermal conductor has a high k (most of metals) while thermal insulators (most of dielectrics) have a low k. Thermal conductivity is considered constant, however, it somewhat increases with temperature. To calculate heat conduction through, say, an electric wire, temperatures at both ends (T_1 and T_2) must be used in equation

$$H = kA\frac{T_1 - T_2}{L}, \tag{3.119}$$

where L is the length of the wire. Quite often, a thermal resistance r is used instead of a thermal conductivity

$$r = \frac{L}{k}, \tag{3.120}$$

then (3.119) can be rewritten as

$$H = A\frac{T_1 - T_2}{r} \tag{3.121}$$

Values of thermal conductivities for some materials are shown in Table A.17.

Figure 3.39 shows an idealized temperature profile within the layers of laminated materials having different thermal conductivities. In the real world, heat transfer through an interface of two adjacent materials may be different from that idealized case. If we join together two materials and observe the heat propagation through the assembly, a temperature profile may look like the one shown in Fig. 3.40a. If the sides of the materials are well insulated, under steady-state conditions, the heat flux must be the same through both materials. The sudden temperature drop at the boundary region, having surface area, a, is the result of a thermal contact resistance. Heat transfer through the assembly can be described as

$$H = \frac{T_1 - T_3}{R_A + R_c + R_B}, \tag{3.122}$$

where R_A and R_B are thermal resistances of two materials and R_c is the contact resistance

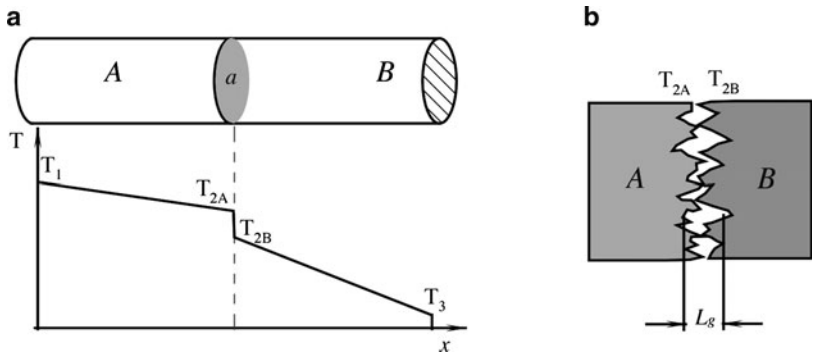

Fig. 3.40 Temperature profile in a joint (**a**) and a microscopic view of a surface contact (**b**)

$$R_{\mathrm{c}} = \frac{1}{h_{\mathrm{c}} a}. \tag{3.123}$$

The quantity h_{c} is called the contact coefficient. This factor can be very important in a number of sensor applications because of many heat-transfer situations, which involve mechanical joining of two materials. Microscopically, the joint may look like the one shown in Fig. 3.40b. No real surface is perfectly smooth, and the actual surface roughness is believed to play a central role in determining the contact resistance. There are two principal contributions to the heat transfer at the joint:

1. The material-to-material conduction through the actual physical contact
2. The conduction through trapped gases (air) in the void spaces created by the rough surfaces

Since thermal conductivity of gases is very small as compared with many solids, the trapped gas creates the most resistance to heat transfer. Then, the contact coefficient can be defined as

$$h_c = \frac{1}{L_g}\left(\frac{a_c}{a}\frac{2k_A k_B}{k_A + k_B} + \frac{a_v}{a}k_f\right), \tag{3.124}$$

where L_{g} is the thickness of the void space, k_{f} is the thermal conductivity of the fluid (for instance, air) filling the void space, a_{c} and a_{v} are areas of the contact and void, respectively, and k_A and k_B are the respective thermal conductivities of the materials. The main problem with this theory is that it is very difficult to determine experimentally areas a_{c} and a_{v}, and distance L_{g}. This analysis, however, allows us to conclude that the contact resistance should increase with a decrease in the ambient gas pressure. On the other hand, contact resistance decreases with an increase in the joint pressure. This is a result of a deformation of the high spots of the contact surface, which leads to enlarging a_{c} and creating a greater contact area between the materials. To decrease the thermal resistance, a dry contact between materials should be avoided. Before joining, surfaces may be coated with

fluid having low thermal resistance. For instance, silicone thermal grease is often used for the purpose.

3.12.2 Thermal Convection

Another way to transfer heat is convection. Convection requires an intermediate agent (fluid: gas or liquid), which takes heat from a warmer body, carries it to a cooler body, releases heat, and then may or may not return back to a warmer body to pick up another portion of heat. Heat transfer from a solid body to a moving agent or within the moving agent is also called convection. Convection may be natural (gravitational) or forced (produced by a mechanism). With the natural convection of air, buoyant forces produced by gravitation act upon air molecules. Warmed up air rises carrying heat away from a warm surface. Cooler air descends toward the warmer object. Forced convection of air is produced by a fan or blower. Forced convection is used in liquid thermostats to maintain the temperature of a device at a predetermined level. The efficiency of a convective heat transfer depends on the rate of media movement, temperature gradient, surface area of an object, and thermal properties of moving medium. An object whose temperature is different from the surroundings will lose (or receive) heat, which can be determined from the Newton's law of cooling, which is governed by equation similar to that of a thermal conduction

$$H = \alpha A(T_1 - T_2), \tag{3.125}$$

where convective coefficient α depends on the fluid's specific heat, viscosity, and a rate of movement. The coefficient is not only gravity dependent, its value changes somewhat with the temperature gradient. For a horizontal plate in air the value of α for the gravitational convection may be estimated from

$$\alpha = 2.49\sqrt[4]{T_1 - T_2}\ \mathrm{W/m^2K}, \tag{3.126}$$

while for a vertical plate it is:

$$\alpha = 1.77\sqrt[4]{T_1 - T_2}\ \mathrm{W/m^2K}. \tag{3.127}$$

It should be noted however that these values are applicable for one side of a plate only, assuming that the plate is a surface of an infinite heat source (that is, its temperature doesn't depend on heat loss) and the surroundings have constant temperature. If volume of air is small, like in the air gap between two surfaces of different temperatures, movement of gaseous molecules becomes very restricted due to viscosity and convective heat transfer becomes insignificant. In these cases, thermal conductivity of air and radiative heat transfer should be considered instead.

3.12.3 Thermal Radiation

It was mentioned above that in any object every atom and every molecule vibrate. The average kinetic energy of vibrating particles is represented by the temperature. Each vibrating atom contains a nucleus and an electronic cloud, which is an orbiting electric charge. According to laws of electrodynamics, a moving electric charge is associated with a variable electric field which produces an alternating magnetic field. In turn, when the magnetic field changes, it results in a coupled with it changing electric field, and so on. Thus, a vibrating particle is a source of electromagnetic field which propagates outwardly with the speed of light and is governed by the laws of optics, that is, the electromagnetic waves can be reflected, filtered, focused, etc. The electromagnetic radiation associated with heat is called thermal radiation. Figure 3.41 shows the total electromagnetic radiation spectrum spreading from γ rays to radio waves. Thermal radiation is predominantly situated in the mid- and far-infrared (IR) spectral ranges.

The wavelength of the radiation directly relates to frequency, v, by means of speed of light c in a particular media:

$$\lambda = \frac{c}{v}. \tag{3.128}$$

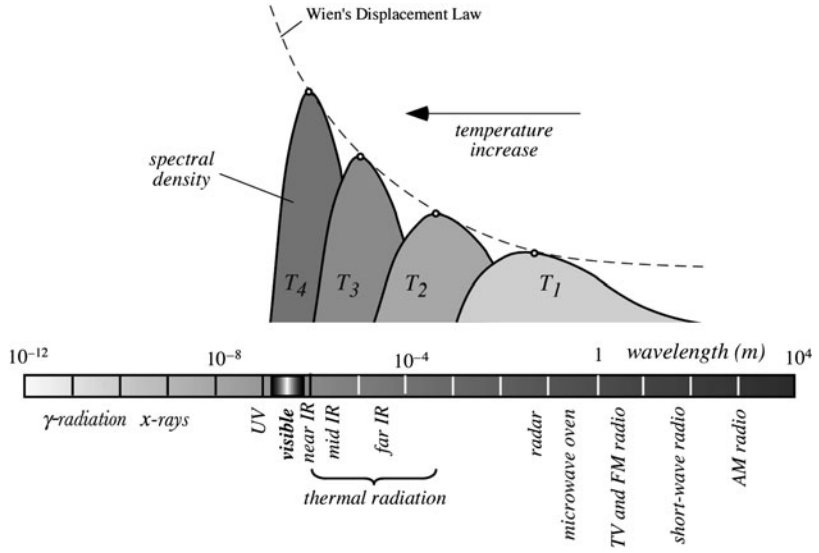

Fig. 3.41 Spectrum of electromagnetic radiation

A relationship between λ and temperature is more complex and is governed by Planck's law, which was discovered in 1901.[23] It establishes radiant flux density W_λ as function of a wavelength λ and absolute temperature T. Radiant flux density is power of electromagnetic radiation per unit of wavelength:

$$W_\lambda = \frac{\varepsilon(\lambda) C_1}{\pi \lambda^5 (e^{C_2/\lambda T} - 1)}, \tag{3.129}$$

where $\varepsilon(\lambda)$ is the emissivity of an object, $C_1 = 3.74 \times 10^{-12}$ Wcm^2 and $C_2 = 1.44$ cmK are constants, and e is the base of natural logarithms. Note that this fundamental equation defines the radiant power at a specific wavelength as function of the object temperature.

Temperature is a result of averaged kinetic energies of an extremely large number of vibrating particles. However, all particles do not vibrate with the same frequency or magnitude. Different permissive frequencies (also wavelengths and energies) are spaced very close to one another, which makes the material capable of radiating of a virtually infinite number of frequencies, spreading from very long to very short wavelengths. Since temperature is a statistical representation of an average kinetic energy, it determines the highest probability for the particles to vibrate with a specific frequency and to have a specific wavelength. This most probable wavelength is established by the Wien's law,[24] which can be found by equating to zero a first derivative of (3.129). The result of the calculation is a wavelength near which most of the radiant power is concentrated:

$$\lambda_m = \frac{2898}{T}, \tag{3.130}$$

where λ_m is in μm and T in K. Wien's law states that the higher the temperature the shorter the wavelength (Fig. 3.41). In view of (3.128), the Wien's law also states that the most probable frequency in the entire spectrum is proportional to the absolute temperature

$$\nu_m = 10^{11} T[\text{Hz}]. \tag{3.131}$$

For instance, at normal room temperature most of the mid- and far-infrared energy is radiated from an object near 30 THz (30×10^{12} Hz). According to the Planck's equation, radiated frequencies and wavelengths depend only on temperature, while the magnitude of radiation also depends on the emissivity $\varepsilon(\lambda)$ of the

[23] In 1918, Max K. E. L. Planck (Germany, Berlin University) was awarded Nobel Prize "in recognition of his services he rendered to the advancement of Physics by his discovery of energy quanta".

[24] In 1911, Wilhelm Wien (Germany, Würtzburg University) was awarded Nobel Prize "for his discoveries regarding the laws governing the radiation of heat".

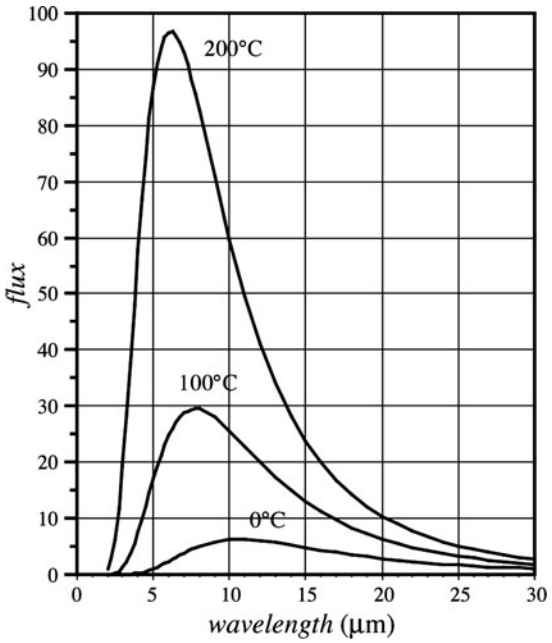

Fig. 3.42 Spectral flux density for three temperatures for an ideal radiator emanating toward infinitely cold space

surface, which we discuss below in detail as it is an important characteristic of thermal radiation sensors.

Figure 3.42 shows the radiant flux density for three different temperatures for the infinitely wide bandwidth (λ_1=0 and λ_2=∞). It is seen, that the radiant energy is distributed over the spectral range highly nonuniformly, with clearly pronounced maximum defined by the Wien's law. A hot object radiates a significant portion of its energy in the visible range, while the power radiated by the cooler objects is concentrated in the near, mid and far infrared (IR) portions of the spectrum.

Theoretically, a thermal radiation bandwidth is infinitely wide. However, when detecting that radiation, properties of the real world sensors must be accounted for. Any sensor is capable of measuring only a limited spectral range (bandwidth) of radiation. In order to determine the total radiated power within a particular bandwidth, (3.129) is integrated within the range limits from λ_1 to λ_2:

$$\Phi_{bo} = \frac{1}{\pi} \int_{\lambda_1}^{\lambda_2} \frac{\varepsilon(\lambda) C_1 \lambda^{-5}}{e^{C_2/\lambda T} - 1} d\lambda, \tag{3.132}$$

Equation (3.132) is quite complex and can't be solved analytically for any particular bandwidth. A solution can be found either numerically or by an approximation. An approximation for a broad bandwidth (when λ_1 and λ_2 embrace over 50% of the total radiated power) is a 4th-order parabola, which is known as the Stefan-Boltzmann law:

$$\Phi_{bo} = A\varepsilon\sigma T^4 \tag{3.133}$$

Here $\sigma = 5.67 \times 10^{-8}$ W/m^{2}K^4 (Stefan-Boltzmann constant), A is the geometry factor, and emissivity ε is assumed to be wavelength independent [37].

3.12.3.1 Emissivity

While wavelengths of the radiated IR light are temperature-dependent, the magnitude of radiation is also a function of the surface property, which is called emissivity, ε. Emissivity is measured on a scale from 0 to 1. It is a ratio of the IR flux that is emanated from a surface to that would be emanated from an ideal emitter (ε=1) having the same temperature. There is a fundamental equation that connects emissivity ε, transparency γ, and reflectivity ρ of an object:

$$\varepsilon + \gamma + \rho = 1 \tag{3.134}$$

In 1860, Kirchhoff had found that emissivity and absorptivity, α, is the same thing. As a result, for an opaque object (γ=0), reflectivity, ρ, and emissivity, ε, are connected by a simple relationship: $\rho = 1 - \varepsilon$.

The Stefan-Boltzmann law specifies the radiant power (flux) that would be emanated from a surface of temperature, T, toward an infinitely cold space (at absolute zero). When thermal radiation is detected by a thermal sensor,[25] the opposite flowing radiation from the sensor toward the object must also be accounted for. A thermal sensor is capable of responding only to a net thermal flux, i.e., flux from the object minus flux from itself toward the object. The surface of the sensor that faces the object has emissivity, ε_s (and, subsequently reflectivity $\rho_s = 1 - \varepsilon_s$). Since the sensor is only partly absorptive, not the entire flux, Φ_{bo}, is absorbed and utilized. One part of it, Φ_{ba}, is absorbed by the sensor while another part, Φ_{br}, is reflected back toward to object[26] (Fig. 3.43). The reflected flux is proportional to the sensor's coefficient of reflectivity

$$\Phi_{br} = -\rho_s \Phi_{bo} = -A\varepsilon(1 - \varepsilon_s)\sigma T^4 \tag{3.135}$$

A negative sign indicates an opposite direction with respect to flux Φ_{bo}. As a result, the net flux originated by the object is

$$\Phi_b = \Phi_{bo} + \Phi_{br} = A\varepsilon\varepsilon_s\sigma T^4. \tag{3.136}$$

[25] Here we discuss the so-called thermal sensors as opposed to quantum sensors that are described in Chapter 13.

[26] This simplified analysis assumes that there are no other objects in the sensor's field of view.

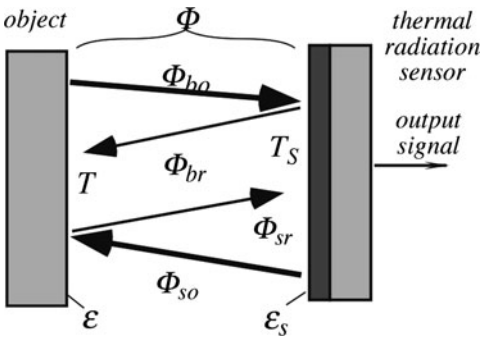

Fig. 3.43 Thermal radiation exchange between an object and a thermal radiation sensor

Depending on its temperature T_s, the sensor's surface radiates its own thermal flux that in a similar way results in the net thermal flux

$$\Phi_s = -A\varepsilon\varepsilon_s\sigma T_S^4. \tag{3.137}$$

Two net fluxes that originate by the object and by the sensor are combined into a final net flux existing between two surfaces

$$\Phi = \Phi_b + \Phi_s = A\varepsilon\varepsilon_s\sigma(T^4 - {T_s}^4). \tag{3.138}$$

This is a mathematical model of a net thermal flux that is converted by a thermal sensor into the output signal. It establishes a connection between thermal power, Φ, absorbed by the sensor, and the absolute temperatures of the object and sensor.

The surface emissivity of a media is function of its dielectric constant and, subsequently, refractive index n. The highest possible emissivity is 1. It is attributed to the so-called *blackbody*, an ideal emitter of electromagnetic radiation. The name implies its appearance at normal room temperatures. If the object is opaque (γ=0) and nonreflective (ρ=0) according to (3.134), it becomes an ideal emitter and absorber of electromagnetic radiation (since ε=α). Thus, a blackbody is an ideal emitter and absorber of light. In practice, a true blackbody does not exist. However it can be approached quite closely. A well-designed blackbody should have emissivity near ε=0.999. A lower emissivity approximately from 0.98 to 0.99 is attributed to the so-called *graybody*.

It should be noted that emissivity is generally wavelength-dependent (Fig. 3.44). For example, a white sheet of paper is very much reflective in the visible spectral range and emits virtually no visible light. However, in the near- and far-infrared spectral ranges its reflectivity is low and emissivity is high (about 0.92), thus making paper a good emitter of far infrared radiation. Polyethylene, which is widely used for fabrication of far infrared lenses, heavily absorbs (emits) in narrow bands around 3.5, 6.8, and 13.5 μm, while being quite transparent (nonemissive) in other bands.

For many practical purposes, emissivity of an opaque material in a relatively narrow spectral range of thermal radiation may be considered constant. For

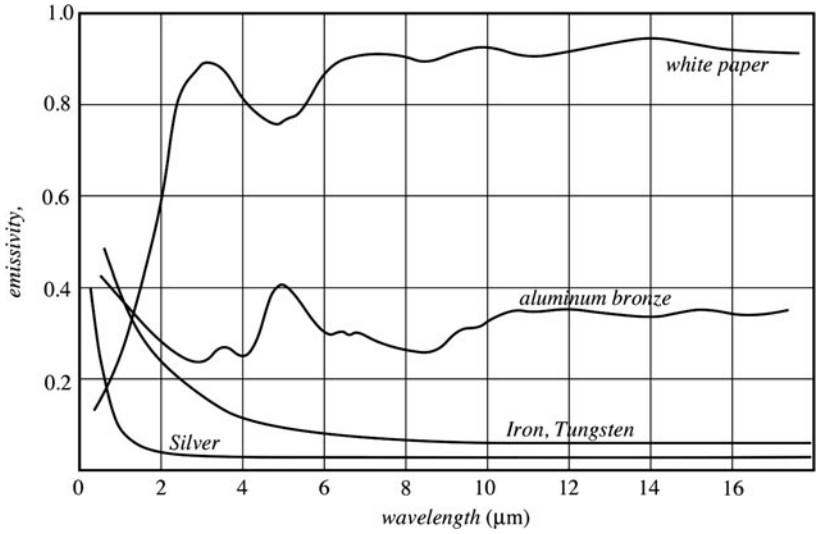

Fig. 3.44 Wavelength dependence of emissivities

precision, measurements, when thermal radiation must be detected with accuracy better than 1%, surface emissivity either must be known, or special methods to reduce effects of emissivity should be employed. One such method is the so-called dual-band IR detectors.[27] The other method takes the benefit of a thermally balanced infrared sensor.[28]

For a nonpolarized mid- and far-infrared light in normal direction, emissivity may be expressed through refractive index by the equation

$$\varepsilon = \frac{4n}{(n+1)^2}. \tag{3.139}$$

All nonmetals are very good diffusive emitters of thermal radiation with a remarkably constant emissivity defined by (3.139) within a solid angle of about $\pm 70^\circ$. Beyond that angle, emissivity begins to decrease rapidly to zero with the angle approaching 90°. Near 90° to normal emissivity is very low. A typical calculated graph of the directional emissivity of nonmetals into air is shown in Fig. 3.45a. It should be emphasized that the above considerations are applicable only to wavelengths in the mid- and far-infrared spectral range and are not true for

[27]Dual band detectors use two narrow spectral ranges to detect the IR flux. Then, by using a ratiometric technique of signal processing, temperature of an object is calculated. During the calculation, emissivity and other multiplicative constants are cancelled out.

[28]In a thermally balanced IR sensor, the sensor's temperature is constantly controlled (warmed up or cooled down) to bring the net thermal flux close to zero. Then, according to (3.138), the emissivities are multiplied by zero and thus their values no longer make any difference.

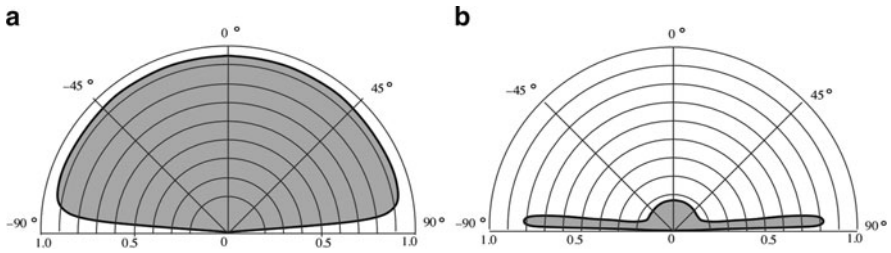

Fig. 3.45 Spatial emissivities for nonmetal (**a**) and a polished metal (**b**)

the visible light, since emissivity of thermal radiation is a result of electromagnetic effects which occur at an appreciable depth below the surface of a dielectric.

Metals behave quite differently. Their emissivities greatly depend on surface finish. Generally, polished metals are poor emitters (good reflectors) within the solid angle of $\pm 70^\circ$ while their emissivity increases at larger angles (Fig. 3.45b). This implies that even a very good metal mirror reflects poorly at angles approaching 90° to normal. Table A.18 in Appendix gives typical emissivities of some materials in a temperature range between 0 and 100°C.

Unlike most solid bodies, gases in many cases are transparent to thermal radiation. When they absorb and emit radiation, they usually do so only in certain narrow spectral bands. Some gases, such as N_2, O_2, and others of nonpolar symmetrical molecular structure, are essentially transparent at low temperatures, while CO_2, H_2O, and various hydrocarbon gases radiate and absorb to an appreciable extent. When infrared light enters a layer of gas, its absorption has an exponential decay profile, governed by Beer's law

$$\frac{\Phi_x}{\Phi_0} = \mathrm{e}^{-\alpha_\lambda x}. \tag{3.140}$$

where Φ_0 is the incident thermal flux, Φ_x is the flux at thickness x, α_λ is the spectral coefficient of absorption (emissivity). The above ratio is called a monochromatic transmissivity γ_λ at a specific wavelength λ. If gas is nonreflecting, then its emissivity at a specific wavelength λ is defined as

$$\varepsilon_\lambda = 1 - \gamma_\lambda = 1 - \mathrm{e}^{-\alpha_\lambda \mathrm{x}}. \tag{3.141}$$

It should be emphasized that since gasses absorb only in narrow bands, emissivity and transmissivity must be specified separately for any particular wavelength. For instance, water vapor is highly absorptive at wavelengths of 1.4, 1.8, and 2.7 μm and is very transparent at 1.6, 2.2 and 4 μm.

Knowing emissivity is essential when an infrared sensor is used for a noncontact temperature measurement (see (3.138)). To calibrate such a noncontact thermometer or verify its accuracy, a laboratory standard source of heat must be constructed. The source must have precisely known emissivity and that emissivity preferably

should approach unity as close as practical. In other words, a blackbody should be used as a laboratory source of a mid- and far-IR radiation. A nonunity emissivity would result in reflection (3.134) that may introduce a significant error in measured infrared flux. There is no known material that has emissivity of one. Thus, a practical way to artificially simulate such a surface is use of the cavity effect, which is the basis for constructing a blackbody.

3.12.3.2 Cavity Effect

An interesting effect develops when electromagnetic radiation is measured from a cavity. For this purpose, a cavity means a void of a generally irregular shape inside a body whose inner wall temperature is uniform over an entire surface (Fig. 3.46a). Emissivity of the cavity opening or aperture (not of the cavity inner surface!) dramatically increases approaching unity at any wavelength, as compared with a flat surface. The cavity effect is especially pronounced when the inner walls of a void have relatively high emissivity. Let us consider a nonmetal surface of a cavity. All nonmetals are diffuse emitters. Also, they are diffuse reflectors. We assume that temperature and emissivity of the cavity walls are homogeneous over the entire area. An ideal surface would emanate from some area a the ideal infrared photon flux $\Phi_0 = a\sigma T_b^4$. However, the real surface has lower emissivity ε_b and, as a result, the flux radiated from that area is smaller: $\Phi_r = \varepsilon_b \Phi_0$. The flux, which is emitted by other parts of the cavity toward area a, is also equal to Φ_r (since the object is thermally homogeneous we may disregard spatial distribution of flux). A substantial portion of that incident flux Φ_r is absorbed by the surface of area a, while a smaller part is diffusely reflected:

$$\Phi_\rho = \rho \Phi_r = (1 - \varepsilon_b)\varepsilon_b \Phi_0 \tag{3.142}$$

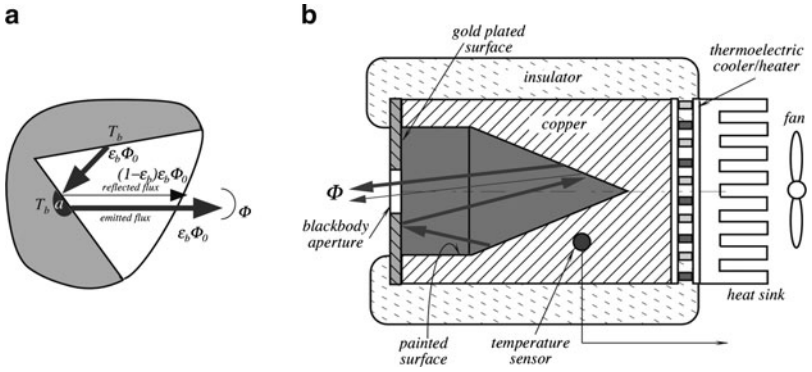

Fig. 3.46 Cavity effect enhances emissivity (**a**); construction of a practical blackbody with a dual cavity surface (**b**)

and the combined radiated and reflected flux from area a toward the cavity aperture is

$$\Phi = \Phi_r + \Phi_\rho = \varepsilon_b \Phi_0 + (1 - \varepsilon_b)\varepsilon_b \Phi_0 = (2 - \varepsilon_b)\varepsilon_b \Phi_0. \tag{3.143}$$

As a result, the effective emissivity of the aperture is expressed as

$$\varepsilon_e = \frac{\Phi}{\Phi_0} = (2 - \varepsilon_b)\varepsilon_b \tag{3.144}$$

It follows from the above that due to just a single reflection, a perceived (effective) emissivity of a cavity is equal to the surface emissivity magnified by a factor of $(2 - \varepsilon_b)$. For example, assuming that the cavity wall is painted by an acrylic paint (emissivity 0.95), the aperture effective emissivity becomes ε_e=(2–0.96)0.96=0.998. Of course, there may be more than one reflection of radiation before it exits the cavity. In other words, the incident on area a flux could already be a result of a combined cavity effect from the reflectance and emittance at other parts of the cavity's surface. The flux intensity will be higher than the originally emanated flux Φ_r.

For a cavity effect to work, the effective emissivity must be attributed to the cavity opening from which radiation escapes. If a sensor is inserted into the cavity too deeply facing its wall directly blocking the reflected rays, the cavity effect will disappear and the emissivity becomes closer to that of the wall surface, which is always lower than unity.

The cavity effect changes a perceived emissivity of a surface and, if not accounted for, may cause error in evaluation of the radiated power. To illustrate this, Fig. 3.47 shows two photographs of a human face - one is taken in visible light and the other in mid infrared. Note that areas at the nostrils appear a little bit brighter (that is warmer). Yet the skin temperature in these spots is the same as

Fig. 3.47 Photographs in visible light and IR thermal radiation that is naturally emanated from the object. Note the brighter (appearing warmers) areas at the wrinkles and skin folds near the nose - a result of the cavity effect. Eyeglasses appear black (cold) because glass in opaque in the mid- and far infrared spectral ranges and does not pass thermal radiation (photo courtesy of Infrared Training Center, www.infraredtraining.com)

nearby. Two wrinkles above the mustache cause the cavity effect, which increases the skin emissivity from an average of 0.96 to a higher value. This enhances intensity of the emanated thermal flux and creates an illusion of a warmer skin.

Design and fabrication of a blackbody is not a trivial task. For a cavity effect to work, a blackbody must have a cavity whose surface area is much larger than the exit aperture, the shape of the cavity must allow for multiple inner reflections before the flux may escape from the aperture, and the cavity wall temperature must be highly uniform all over its entire surface. Figure 3.46b shows an efficient way of fabricating a blackbody [38] whose emissivity exceeds 0.999. A cavity body is fabricated of a solid copper or aluminum with the cavity of any practical shape, while an inversed cone is preferable. An imbedded temperature sensor and a thermoelectric heater/cooler with a control circuit (not shown) form a thermostat that maintains temperature of the cavity on a preset level. That may be above or below the ambient temperature. The inner portion of a cavity should be painted with organic paint. The visible color of paint is not important as there is no correlation between the paint reflectivity in a visible range (which determines visible color) and its emissivity in the infrared spectral range. The most troublesome portion of a cavity is located near the aperture, since it is very difficult to assure temperature of the left side of the blackbody, as in Fig. 3.46b, to be independent of the ambient temperature and be equal to the rest of the cavity walls. To minimize effects of the ambient temperature and increase a virtual cavity size, the inners surface of the front wall around the cavity is highly polished and gold plated. Thus, the front side of the cavity has very low emissivity and thus its temperature is not that critical. Besides, the gold surface reflects rays emitted by the right-side parts of the cavity walls that have high emissivity and thus enhances the cavity effect. The entire metal body is covered with a thermally insulating layer. It should be stressed again that the blackbody surface is the virtual surface of an aperture, which reality is a void.

3.13 Light

Light is a very efficient form of energy for sensing a great variety of stimuli. Among many others, these include distance, motion, temperature, chemical composition and many others. Light has an electromagnetic nature. It may be considered a propagation of either quanta of energy or electromagnetic waves. This confusing duality nowadays is well explained by quantum electrodynamics [39]. Different portions of the wave frequency spectrum are given special names, for example: ultraviolet (UV), visible, near, mid- and far-infrared (IR), microwaves, radio waves, etc. The name “light” was arbitrarily given to electromagnetic radiation which occupies wavelengths from approximately 0.1 to 100 μm. Light below the shortest wavelength that we can see (violet) is called ultraviolet and farther than the longest that we can see (red) is called infrared. The infrared range is arbitrarily subdivided into three regions: near-infrared (from about 0.9 to 1.5 mμ), mid-infrared (1.5 to 5 μm), and far-infrared (5 to 100 μm).

Different portions of the radiation spectrum are studied by separate branches of physics and employed by different branches of engineering. An entire electromagnetic spectrum is represented in Fig. 3.41. It spreads from γ rays (the shortest) to radiowaves (the longest). In this section, we will briefly review those properties of light which are mostly concerned with the visible and near infrared portions of the electromagnetic spectrum. Thermal radiation (mid- and far-IR regions) is covered in Sect. 3.12.3.

The velocity of light c_0 in vacuum is independent of wavelengths and can be expressed through $\mu_0 = 4\pi \times 10^{-7} \frac{\text{henry}}{\text{m}}$ and $\varepsilon_0 = 8.854 \times 10^{-12} \frac{\text{farad}}{\text{m}}$, which are the magnetic and electric permitivities of free space:

$$c_0 = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 299,792,458.7 \pm 1.1 \frac{\text{m}}{\text{s}}. \tag{3.145}$$

The frequency of light waves in vacuum or any particular medium relates to its wavelength λ by (3.128), which we rewrite here as

$$v = \frac{c}{\lambda}, \tag{3.146}$$

where c is the speed of light in a medium.

The energy of a photon relates to its frequency as

$$E = hv, \tag{3.147}$$

where $h = 6.63 \times 10^{-34}$ J·s (4.13×10^{-15} eV·s) is the Planck's constant. The photon energy E is measured in 1.602×10^{-19} J = 1 eV (electron-volt).

The UV and visible photons carry relatively large energy and are not difficult to detect. However, when the wavelength increases and moves to an infrared portion of the spectrum, the detection becomes more and more difficult. It follows from (3.147) that energy of quanta drops for lower frequencies. A near-infrared photon having a wavelength of 1 μm has the energy of 1.24 eV. Hence, an optical quantum detector operating in the range of 1 μm must be capable of responding to that level of energy. If we keep moving even further, toward the mid- and far-infrared spectral ranges, we deal with smaller energies. Human skin (at 37°C) radiates near and far infrared photons with energies near 0.13 eV, which is an order of magnitude lower than red light, making them much more difficult to detect. This is the reason why low energy radiation is often detected by thermal detectors rather than quantum detectors.

3.13.1 Light Polarization

The electromagnetic wave (now we ignore the quantum properties of light) has the additional characteristic that is polarization (more specifically, plane polarization). This means that the alternating electric field vectors are parallel to each other for all points in the wave. The magnetic field vectors are also parallel to each other, but in

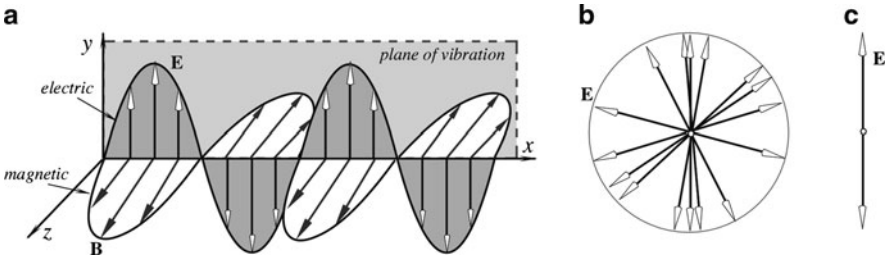

Fig. 3.48 Traveling electromagnetic wave has electric and magnetic field vectors (**a**); unpolarized electric field viewed along the *x*-axis (magnetic vectors are not shown but they are always there) (**b**); vertically polarized electric field (**c**)

dealing with the polarization issues related to sensor technologies, we focus our attention on the electric field, to which most detectors of the electromagnetic radiation are sensitive. Figure 3.48a shows the polarization feature. The wave in the picture is traveling in the *x*-direction. It is said the wave to be polarized in the *y*-direction because the electric field vectors are all parallel to this axis. The plane defined by the direction of propagation (the *x*-axis) and the direction of polarization (the *y*-axis) is called the plane of vibration. In a polarized light, there are no other directions for the field vectors.

Figure 3.48b shows a randomly polarized light, which is the type of light that is produced by the sun and various incandescent light sources, however, the emerging beam in most laser configurations is polarized. If unpolarized light passes through a polarization filter (Polaroid), only specific planes can pass through and the output electric filed will be as shown in Fig. 3.48c. The polarization filter transmits only those wavetrain components whose electric vectors vibrate parallel to the filter direction and absorbs those that vibrate at right angles to this direction. The emerging light will be polarized according to the filter orientation. This polarizing direction in the filter is established during the manufacturing process by embedding certain long-chain molecules in a flexible plastic sheet and then stretching the sheet so that the molecules are aligned in parallel to each other. The polarizing filters are most widely used in the liquid crystal displays (LCD) and in many optical sensors that are described in the corresponding chapters of this book.

3.13.2 Light Scattering

Scattering is an electromagnetic phenomenon where light is forced to deviate from a straight path by one or more localized nonuniformities in the medium [30]. Examples of nonuniformities are smoke particles, dust, bacteria, water droplets, and gaseous molecules. When a particle is larger than the wavelength of incident light, a particle that happens to be in the light path serves as a reflector (Fig. 3.49b). The reflection is governed by general laws of reflection as described above. Reflections that undergo scattering are often called diffuse reflections, or unscattered reflections, or specular (mirror-like) reflections.

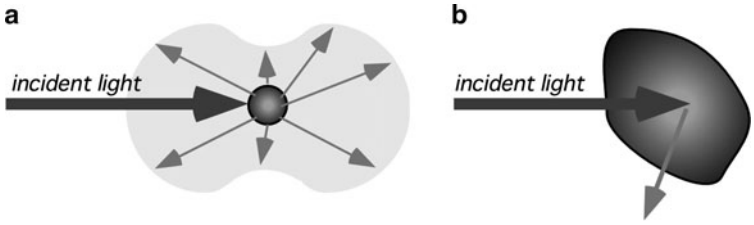

Fig. 3.49 Scattering of light from small (**a**) and large (**b**) particles

Smaller particles exhibit a different scattering. It is typical for particles that are at least ten times smaller than the wavelength of light. In a simplified way, the scattering mechanism by a small particle can be explained as absorption of the light energy and re-emitting it in all directions (Fig. 3.49a). A scattering theory studies electromagnetic radiation (light) scattered by a small spherical volume of variant refractive index, such as a particle, bubble, droplet, or even a density fluctuation. This effect was first modeled by Lord Rayleigh, from whom it gets its name. For a very small particle, the exact shape of the scattering particle is usually not very significant and can often be treated as a sphere of equivalent volume.

The inherent scattering that light undergoes passing through a pure gas is due to microscopic density fluctuations as the gas molecules move around, which are normally small enough in scale for Rayleigh's model to apply. Scattering depends on size of the particle or irregularity, wavelength of light, and angle between the scattered and incident lights. This scattering mechanism is the primary cause of the blue color of the Earth's sky on a clear day, as the shorter blue wavelengths of sunlight passing overhead are more strongly scattered than the longer red wavelengths at angles significantly deviating from the direction of the light beam coming from the Sun. However, at a sunset, at angles closer to the direction to the Sun, the sky appears orange and red because of a stronger scattering of longer (red) wavelengths. At night, the sky appears black because the sun light rays pass above the atmosphere and thus are not scattered toward the earth observer. Therefore, a sensor that employs scattering by small particles, can operate either on a principle of measuring the light intensity or a shift in the scattered light spectrum.

Light scattering is a phenomenon that can be used for detecting small impurities in gases and liquid and for sensing concentration of particles in a fluid. Examples of the applications are a smoke detector and air cleanliness monitor, which senses presence of dust.

3.14 Dynamic Models of Sensor Elements

To determine a sensor's dynamic response a variable stimulus should be applied to its input while observing the output values (see Sect. 2.18). Generally, a test stimulus may have any shape or form, which should be selected depending on a

practical need. For instance, while determining a natural frequency of an accelerometer, sinusoidal vibrations of different frequencies are the best. On the other hand, for a thermistor probe, a step-function of temperature would be preferable. In many other cases, a step or square-pulse input stimulus is often employed. The reason is that they have a theoretically infinite frequency spectrum. That is, the sensor simultaneously can be tested at all frequencies.

Mathematical modeling of a sensor is a powerful tool in assessing its performance. The modeling may address two issues: static and dynamic. The models usually deal with the sensor's transfer function as it is defined in Chap. 2. Here we briefly outline how sensors can be evaluated dynamically. The dynamic models may have several independent variables, however, one of them must be time. The resulting model is referred to as a lumped parameter model. In this section, mathematical models are formed by applying physical laws to some simple lumped parameter sensor elements. In other words, for the analysis, a sensor is divided into simple elements and each element is considered separately. However, once the equations describing the elements have been formulated, individual elements can be recombined to yield the mathematical model of the original sensor. The treatment is intended not to be exhaustive, but rather to introduce the topic.

3.14.1 Mechanical Elements

Dynamic mechanical elements are made of masses, or inertias, which have attached springs and dampers. Often the damping is viscous, and for the rectilinear motion the retaining force is proportional to velocity. Similarly, for the rotational motion, the retaining force is proportional to angular velocity. Also, the force, or torque, exerted by a spring, or shaft, is usually proportional to displacement. The various elements and their governing equations are summarized in Table 3.4.

Table 3.4 Mechanical, thermal, and electrical analogies

Mechanical	Thermal	Electrical	
Mass M $F = M\frac{d(v)}{dt}$	Capacitance C $Q = C\frac{dT}{dt}$	Inductor L $V = L\frac{di}{dt}$	Capacitor $i = C\frac{dV}{dt}$
Spring k $F = k\int vdt$	Capacitance C $T = \frac{1}{C}\int Qdt$	Capacitor C $V = \frac{1}{C}\int idt$	Inductor L $i = \frac{1}{L}\int Vdt$
Damper b $F = bv$	Resistance R $Q = \frac{1}{R}(T_2 - T_1)$	Resistor R $V = Ri$	Resistor R $i = \frac{1}{R}V$

One of the simplest methods of producing the equations of motion is to isolate each mass or inertia and to consider it as a free body. It is then assumed that each of the free bodies is displaced from the equilibrium position, and the forces or torques acting on the body then drive it back to its equilibrium position. Newton's second law of motion can then be applied to each body to yield the required equation of motion.

For a rectilinear system Newton's second law indicates that for a consistent system of units *the sum of forces equals to the mass times the acceleration*. In the SI system of units, force is measured in newtons (N), mass in kilograms (kg), and acceleration in meters per second squared (m/s^2).

For a rotational system, Newton's law becomes: *the sum of the moments equals the moment of inertia times the angular acceleration*. The moment, or torque, has units of newton-meters (Nm), the inertia units of kilogram per meter squared (kg/m^2) and the angular acceleration units of radians per second squared (rad/s^2).

Let us consider a monoaxial accelerometer, which consists of an inertia element whose movement may be transformed into an electric signal. The mechanism of conversion may be, for instance, piezoelectric. Figure 3.50a shows a general mechanical structure of such an accelerometer. Mass M is supported by a spring having stiffness k and the mass movement is damped by a damping element with a coefficient b. Mass may be displaced with respect to the accelerometer housing only in the horizontal direction. During operation, the accelerometer case is subjected to acceleration d^2y/dt^2, and the output signal is proportional to the deflection x_0 of the mass M.

Since the accelerometer mass M is constrained to linear motion, the system has one degree of freedom. Giving the mass M a displacement x from its equilibrium position produces the free-body diagram shown in Fig. 3.50b. Note that x_0 is equal to x plus some fixed displacement. Applying Newton's second law of motion gives

$$Mf = -kx - b\frac{dx}{dt}, \tag{3.148}$$

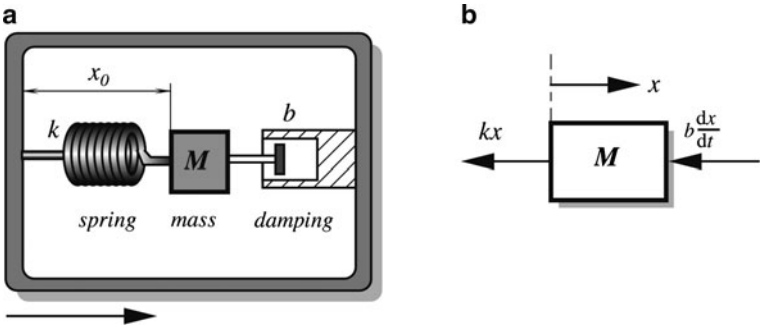

Fig. 3.50 Mechanical model of an accelerometer (**a**) and a free-body diagram of mass (**b**)

where f is the acceleration of the mass relative to the Earth and is given by

$$f = \frac{d^2x}{dt^2} - \frac{d^2y}{dt^2}. \tag{3.149}$$

Substituting for f gives the required equation of motion as:

$$M\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = M\frac{d^2y}{dt^2}. \tag{3.150}$$

Note that each term in the above equation has units of newtons (N). The differential equation (3.150) is of a second order, which means that the accelerometer output signal may have the oscillating shape. By selecting an appropriate damping coefficient b the output signal may be brought to a critically damped state which, in most cases, is a desirable response.

3.14.2 Thermal Elements

Thermal elements include such things as heat sinks, heating and refrigeration elements, insulators, heat reflectors, and absorbers. If heat is of a concern, a sensor should be regarded as a component of a larger device. In other words, heat conduction through the housing and the mounting elements, air convection and radiative heat exchange with other objects should not be discounted (see discussion is Sect. 16.1).

Heat may be transferred by three mechanisms: conduction, natural and forced convection, and thermal radiation (Sect. 3.12). For simple lumped parameter models, the first law of thermodynamics may be used to determine the temperature changes in a body. The rate of change of a body's internal energy is equal to the flow of heat into the body less the flow of heat out of the body, very much like fluid moves through pipes into and out of a tank. This balance may be expressed as

$$C\frac{\mathrm{d}T}{\mathrm{d}t} = \Delta Q, \tag{3.151}$$

where $C{=}Mc$ is the thermal capacity of a body (J/K), T is the temperature (K), ΔQ is the heat flow rate (W), M is the mass of the body (kg), and c is the specific heat of the material (J/kg·K). The heat flow rate through a body is a function of the thermal resistance of the body. This is normally assumed to be linear, and therefore

$$\Delta Q = \frac{T_1 - T_2}{r}, \tag{3.152}$$

where r is the thermal resistance (K/W) and T_1–T_2 is a temperature gradient across the element, where heat conduction is considered.

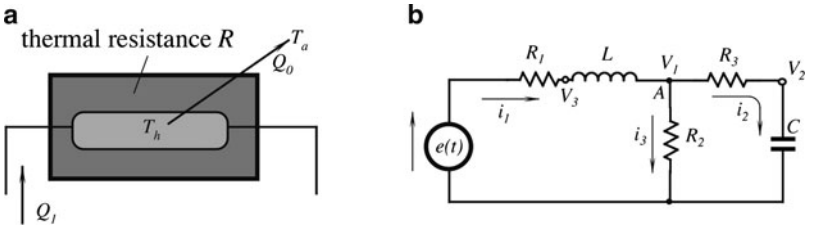

Fig. 3.51 Thermal model of a heating element (**a**); electrical circuit diagram with resistive, capacitive and inductive components (**b**)

For the illustration, we analyze a heating element (Fig. 3.51a) having temperature T_h. The element is coated with insulation. The temperature of the surrounding air is T_a. The value Q_1 is the rate of heat supply to the element, and Q_0 is the rate of heat loss. From (3.151)

$$C\frac{dT_h}{dt} = Q_1 - Q_0, \tag{3.153}$$

but, from (3.152)

$$Q_0 = \frac{T_h - T_a}{r}. \tag{3.154}$$

and, in the result, we obtain a differential equation

$$\frac{dT_h}{dt} + \frac{T_h}{rC} = \frac{Q_1}{C} + \frac{T_a}{rC}, \tag{3.155}$$

This is a first-order differential equation, which is typical for thermal systems. A thermal element, if not part of a control system with a feedback loop, is inherently stable. A response of a simple thermal element may be characterized by a thermal time constant which is a product of thermal capacity and thermal resistance: $\tau_T = Cr$. The time constant is measured in units of time (s) and, for a passively cooling element, is equal to time which takes to reach about 37% of the initial temperature gradient.

3.14.3 Electrical Elements

There are three basic electrical elements: the capacitor, the inductor, and the resistor. Again, the governing equation describing the idealized elements is given in Table 3.4. For the idealized elements, the equations describing the sensor's behavior may be obtained from the Kirchhoff's laws, which directly follow from the law of conservation of energy:

Kirchhoff's 1st law: The total current flowing toward a junction is equal to the total current flowing from that junction, i.e., the algebraic sum of the currents flowing through a junction is zero.

Kirchhoff's 2nd law: In a closed circuit, the algebraic sum of the voltages across each part of the circuit is equal to the applied e.m.f.

Let us assume that we have a sensor whose elements may be represented by a circuit shown in Fig. 3.51b. To find the circuit equation, we will use the 1st Kirchhoff's law, which sometimes is called Kirchhoff's current law. For the node, A

$$i_1 - i_2 - i_3 = 0, \tag{3.156}$$

and for each current

$$\begin{aligned} i_1 &= \frac{e - V_3}{R_1} = \frac{1}{L}\int (V_3 - V_1)dt \\ i_2 &= \frac{V_1 - V_2}{R_3} = C\frac{dV_2}{dt} \\ i_3 &= \frac{V_1}{R_2} \end{aligned} \tag{3.157}$$

When these expressions are substituted into (3.156), the resulting equation becomes

$$\frac{V_3}{R_1} + \frac{V_1 - V_2}{R_3} + 2\frac{V_1}{R_2} + C\frac{dV_2}{dt} - \frac{1}{L}\int (V_3 - V_1)dt = \frac{e}{R_1}. \tag{3.158}$$

In the above equation, e/R_1 is the forcing input, and the measurable outputs are V_1, V_2, and V_3. To produce the above equation, three variables i_1, i_2, and i_3 have to be specified and three equations of motion derived. By applying (3.156)) of constrain $i_1 - i_2 - i_3 = 0$ it has been possible to condense all three equations of motion into a single expression. Note that each element in this expression has a unit of current (A).

3.14.4 Analogies

Above, we considered mechanical, thermal, and electrical elements separately. However, the dynamic behavior of these systems is analogous. It is possible, for example, to take mechanical elements or thermal components, convert them into an equivalent electric circuit and analyze the circuit using Kirchhoff's laws. Table 3.4 gives the various lumped parameters for mechanical, thermal, and electrical circuits, together with their governing equations. For the mechanical components, Newton's second law was used and for thermal we apply Newton's law of cooling.

In the first column are the linear mechanical elements and their equations in terms of force (F). In the second column are the linear thermal elements and their equations in terms of heat (Q). In the third and fourth columns are electrical analogies (capacitor, inductor, and resistor) in terms of voltage and current (V and i). These analogies may be quite useful in a practical assessment of a sensor and for the analysis of its mechanical and thermal interface with the object and the environment.

References

1. Halliday D, Resnick R (1986) Fundamentals of physics, 2nd edn. Wiley, New York
2. Constantinos JM, Stergios P, Euan KB (2008) Dalton Trans 1809
3. Crotzer FR Method for manufacturing hygristors. U.S. Patent No. 5,273,777
4. Meissner A (1927) Über piezoelectrische Krystalle bei Hochfrequenz. Z Tech Phys 8:74
5. Neubert HKP (1975) Instrument transducers. An introduction to their performance and design, 2nd edn. Clarendon, Oxford
6. Radice PF (1982) Corona discharge poling process. U.S. Patent No. 4,365, 283
7. Southgate PD (1976) Appl Phys Lett 28:250
8. Jaffe B, Cook WR, Jaffe H (1971) Piezoelectric ceramics. Academic Press, London
9. Mason WP (1950) Piezoelectric crystals and their application to ultrasonics. Van Nostrand, New York
10. Megaw HD (1957) Ferroelectricity in crystals. Methuen, London
11. Oikawa A, Toda K (1976) Preparation of Pb(Zr,Ti)O_3 thin films by an electron beam evaporation technique. Appl Phys Lett 29:491
12. Okada A (1977) Some electrical and optical properties of ferroelectric lead-zirconite-lead-titanate thin films. J Appl Phys 48:2905
13. Castelano RN, Feinstein LG (1979) Ion-beam deposition of thin films of ferroelectric lead-zirconite-titanate (PZT). J Appl Phys 50:4406
14. Adachi H et al (1986) Ferroelectric (Pb, La)(Zr, Ti)O_3 epitaxial thin films on sapphire grown by RF-planar magnetron sputtering. J Appl Phys 60:736
15. Ogawa T, Senda S, Kasanami T (1989) Preparation of ferroelectric thin films by RF sputtering. J Appl Phys 28:11–14
16. Roy D, Krupanidhi SB, Dougherty J (1991) Excimer laser ablated lead zirconite titanate thin films. J Appl Phys 69:1
17. Yi G, Wu Z, Sayer M (1989) Preparation of PZT thin film by sol-gel processing: electrical, optical, and electro-optic properties. J Appl Phys 64
18. Tamura M, Yamaguchi T, Oyaba T, Yoshimi T (1975) J Audio Eng Soc 23:31
19. Elliason S (1984) Electronic properties of piezoelectric polymers. Report TRITA-FYS 6665 from Dept. of Applied Physics, The Royal Inst. of Techn., S-100 44 Stockholm, Sweden
20. Piezo Film Sensors Technical Manual (April 1999) Measurement Specialties, Inc. www.msiusa.com
21. Kawai H (1969) The Piezoelectricity of poly (vinylidene fluoride). Jpn J Appl Phys 8:975–976
22. Meixner H, Mader G, Kleinschmidt P (1986) Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). Siemens Forsch.-u. Entwicl. Ber. Bd. 15(3): 105–114
23. Ye C, Tamagawa T, Polla DL (1991) Pyroelectric PbTiO3 thin films for microsensor applications. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of technical papers, pp 904–907, ©IEEE

24. Beer AC (1963) Galvanomagnetic effect in semiconductors. In: Seitz F, Turnbull D (eds) Supplement to solid state physics. Academic Press, New York
25. Putlye EH (1960) The Hall effect and related phenomena. In: Hogarth (ed) Semiconductor Monographs. Butterworths, London
26. Sprague Hall Effect and Optoelectronic Sensors. Data Book SN-500, 1987
27. Williams J (1990) Thermocouple measurement, AN28, Linear applications handbook, © Linear Technology Corp
28. Seebeck T Dr. Magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz. Abhaandulgen der Preussischen Akademic der Wissenschaften, pp 265–373, 1822–1823
29. Benedict RP (1984) Fundamentals of temperature, pressure, and flow measurements, 3rd edn., Wiley, New York
30. Stover JC (1995) Optical scattering: measurement and analysis. SPIE Optical Engineering Press
31. LeChatelier H (1962) Copt. Tend., 102, 1886,29. In: MacDonald DKC (ed) Thermoelectricity: an introduction to the principles. Wiley, New York
32. Carter EF (ed) (1966) Dictionary of inventions and discoveries. In: Muller F (ed) Crane, Russak, New York
33. Peltier JCA Investigation of the heat developed by electric currents in homogeneous materials and at the junction of two different conductors. Ann Phys Chem 56(2nd ser.):1834
34. Thomson W (May 1854) On the thermal effects of electric currents in unequal heated conductors. Proc Royal Soc VII
35. Manual on the use of thermocouples in temperature measurement. ASTM Publication Code Number 04-470020-40, ©ASTM, Philadelphia, 1981
36. Doebelin EO (1990) Measurement systems: application and design, 4th edn. McGraw-Hill, New York
37. Holman JP (1972) Heat transfer, 3rd edn. McGraw-Hill, New York
38. Fraden J (2002) Blackbody cavity for calibration of infrared thermometers. U.S. Patent No. 6447160
39. Feynman RP (2006) QED The strange theory of light and matter. Princeton University Press, Princeton
40. Kleinschmidt P (1984) Piezo- und pyroelektrische Effekte. In: Heywang W (ed) *Sensorik*. Kap. 6: Springer, Berlin
41. Semiconductor sensors (1988) Data handbook. Philips Export B.V
42. Thompson S (1989) Control systems. Engineering & design. Longman Scientific & Technical, Essex, England

Chapter 4
Optical Components of Sensors

Where the telescope ends, the microscope begins.
Which of the two has the grander view?
–Victor Hugo

Light phenomena such as reflection, refraction, absorption, interference, polarization, and speed are the powerful utensils in a sensor designer's toolbox. Optical components help to manipulate light in many ways. In this chapter, we discuss these components from the standpoint of geometrical optics. When using geometrical optics, we omit the properties of light, which are better described by quantum mechanics and quantum electrodynamics. We will ignore not only the quantum properties of light but the wave properties as well. We consider light as a moving front or a ray, which is perpendicular (normal) to that front. To do so, we should not discuss any optical elements whose dimensions are too small when compared with the wavelength. For example, if a glass window is impregnated with small particles of submicron sizes, we should completely ignore them for any geometrical calculations from the near infrared to longer wavelengths. Another example is diffractive grating. Its operation cannot be described by the methods of geometrical optics. In such cases, the methods of quantum electrodynamics (QED) need to be used. Just briefly we will address the emerging field of nano-optics. In this chapter, we summarize those optical elements that are most applicable for the sensor design. For more detailed discussion of geometrical optics, we refer the reader to special texts, for example [1,15,17].

Before light can be manipulated, first we need to have the light generated. There are several ways to produce light. Some sources of light are natural and exist without our will or effort, while some must be incorporated into a measurement device. The natural sources of light include celestial objects such as sun, moon, and stars. Also, natural sources of light in the mid- and far-infrared spectral ranges include all material objects that radiate thermal energy depending on their temperatures, as it was covered in Chapter 3. These include fire, exothermic chemical reactions, living organisms, and other natural sources whose temperatures are different from their surroundings and whose thermal radiation can be selectively

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_4, © Springer Science+Business Media, LLC 2010

detected by optical devices. The man-made sources of light include filaments in the electric bulbs, light emitting diodes (LED), gas discharge lamps, lasers, laser diodes, heaters, etc.

After light is generated, it can be manipulated in many ways. Figure 4.1 shows several examples of the manipulation of light in sensors. Most of these methods involve changing the direction of light, while some use a selective blocking of certain wavelengths. The latter is called filtering (filter in Fig. 4.1a). The light direction can be changed by the use of a physical effect of reflection with the help of mirrors, diffractive gratings, optical wave guides, and fibers. Also, the light direction can be changed by the refraction with the help of lenses, prisms, windows, chemical solutions, crystals, organic materials, and biological objects. While passing through these objects, properties of light may be modified (modulated) by a measured stimulus. Then, the task of a sensor designer is to arrange a conversion

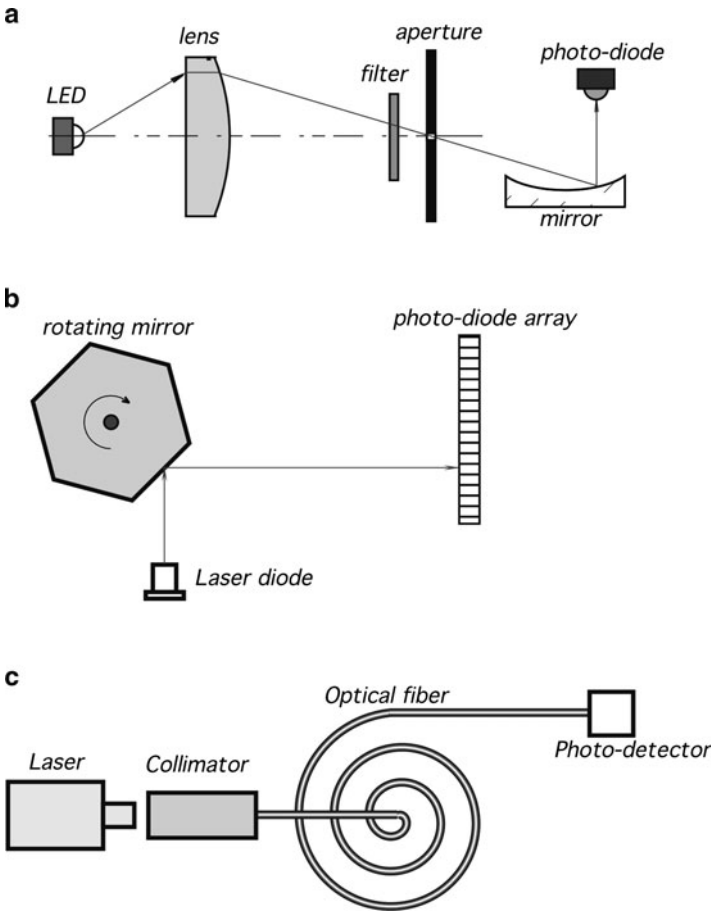

Fig. 4.1 Examples of optical systems that use refraction (**a**) and reflection (**a**, **b**, **c**)

of a such modulation into electrical signals that can be related to the stimulus. What can be modulated in light? The intensity, direction of propagation, polarization, spectral contents of light beam - all these can be modified, and even the speed of light and the phase of its wave can be changed. When developing sensors, one may be concerned with either radiometry or photometry. The former deals with the light power and its manipulation, while the latter is about illumination and its control.

4.1 Radiometry

Let us consider light traveling through a three-layer material. All layers are made of different substances called media. Figure 4.2 shows what happens to a ray of light, which travels from the first medium into a flat plate of a second medium, and then to a third medium. Examples of the media are air, glass, and liquid. Part of the incident light is reflected from a planar boundary between the first and second media according to the law of reflection, which historically is attributed to Heron of Alexandria (first century AD) who noticed that angle of incidence equals angle of reflection

$$\Theta_1 = \Theta'_1 \tag{4.1}$$

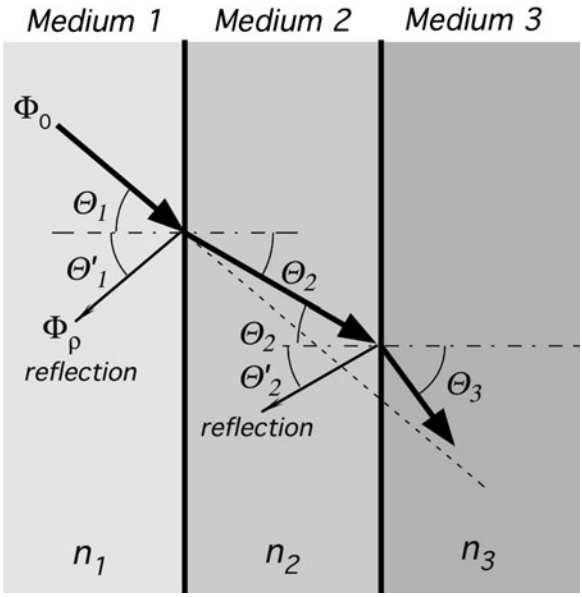

Fig. 4.2 Light passing through materials with different refractive indices

This could be restated by saying that reflected light takes the shortest path or the shortest time to travel between two points. The latter can be derived from the Fermat's principle. This mirror-like reflection is called a specular reflection. Reflection not necessarily should be specular as defined by (4.1). When light strikes a rough or granular boundary between two media, it bounces off in all directions due to the microscopic irregularities of the interface. This is called diffuse reflection. The exact form of reflection depends on the structure of the surface.

Part of light flux enters the plate (medium 2) at a different angle. The new angle Θ_2 is governed by the refraction law, which was discovered in 1621 by Willebrord Snell (1580–1626) and is known as Snell's law:

$$n_1 \sin \Theta_1 = n_2 \sin \Theta_2, \tag{4.2}$$

where n_1 and n_2 are the indices of refraction of two media.

In any medium, light moves slower than in vacuum. An *index of refraction* is a ratio of velocity of light in vacuum, c_0, to that in a medium, c

$$n = \frac{c_0}{c}, \tag{4.3}$$

Since $c < c_0$, the refractive index of a medium is always more than unity. The velocity of light in a medium directly relates to a dielectric constant ε_r of a medium, which subsequently determines the refractive index:

$$n = \sqrt{\varepsilon_r} \tag{4.4}$$

Generally, n is function of a wavelength. A wavelength dependence of index of refraction is manifested in a prism, which was used by Sir Isaac Newton in his experiments with the light spectrum. In the visible range, the index of refraction n is often specified at a wavelength of 0.58756 μm, the yellow-orange helium line. Indices of refraction for some materials are presented in Table A.19 in Appendix.

A refractive index dependence of wavelengths is called *dispersion*. The change in n with the wavelength is usually very gradual, and often negligible, unless the wavelength approaches a region where the material is not transparent. Figure 4.3 shows transparency curves of some optical materials employed in various optical sensors.

A portion of light flux reflected from a boundary at angle Θ'_1 depends on light velocities in two adjacent media. Amount of reflected flux Φ_ρ relates to incident flux Φ_0 through the *coefficient of reflection* ρ, which can be expressed by means of refractive indices

$$\rho = \frac{\Phi_\rho}{\Phi_0} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2. \tag{4.5}$$

Eqs. (3.139) and (4.5) indicate that both the reflection and the absorption (emissivity) depend solely on the refractive index of a material at a particular wavelength.

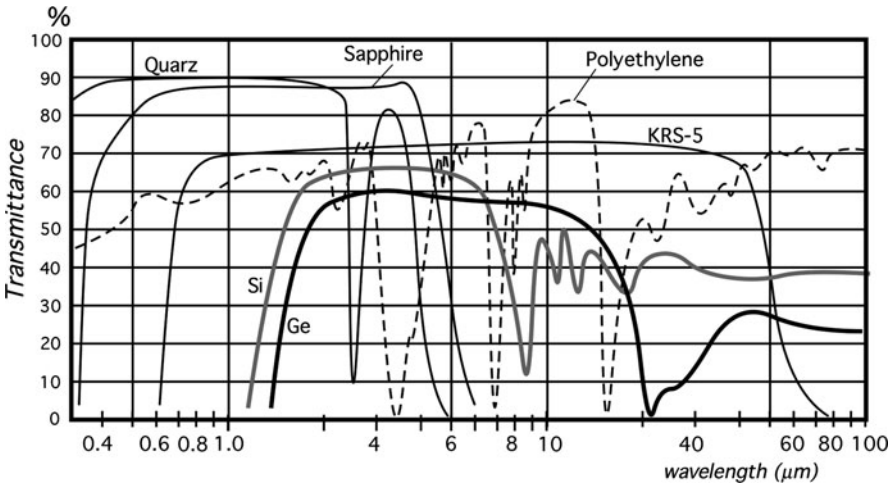

Fig. 4.3 Transparency characteristics for various optical materials

If the light flux enters from air into an object having refractive index n (4.5) is simplified as

$$\rho = \left(\frac{n-1}{n+1}\right)^2, \tag{4.6}$$

Before light exits the second medium (Fig. 4.2) and enters the third medium having refractive index n_3, another part of it is reflected internally from the second boundary between the n_2 and n_3 media at angle Θ'_2. The remaining portion of light exits at angle Θ_3, which is also governed by the Snell law. If media 1 and 3 are the same (for instance, air) at both sides of the plate, then $n_1 = n_3$ and $\Theta_1 = \Theta_3$. This case is illustrated in Fig. 4.4. It follows from (4.5) that coefficients of reflection are the same for light striking a boundary from either direction - approaching from the higher or lower index of refraction.

A combined coefficient of two reflections from both surfaces of a plate can be found from a simplified formula

$$\rho_2 \approx \rho_1(2 - \rho_1) \tag{4.7}$$

where ρ_1 is the reflective coefficient from one surface. In reality, the light reflected from the second boundary is reflected again from the first boundary back to the second boundary, and so on. Thus, assuming that there is no absorption in the material, the total reflective loss within the plate can be calculated through the refractive index of the material

$$\rho_2 = 1 - \frac{2n}{n^2+1} \tag{4.8}$$

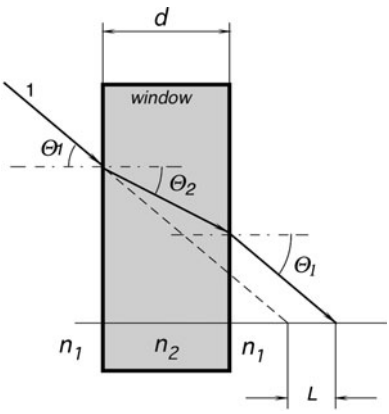

Fig. 4.4 Light passing through an optical plate

Reflection increases for higher differences in refractive indices. For instance, if visible light travels without absorption from air through a heavy flint glass plate, two reflectances result in loss of about 11%, while for the air-germanium-air interfaces (in the far infrared spectral range) the reflective loss is about 59%. To reduce losses, optical materials are often given antireflective coatings (ARC), which have refractive indices and thickness geared to specific wavelengths.

The radiant energy balance (3.134) should be modified to account for two reflections in an optical material:

$$\rho_2 + \alpha + \gamma = 1, \tag{4.9}$$

where α is a coefficient of absorption and γ is a coefficient of transmittance. In a transparency region, $\alpha \approx 0$, therefore, transmittance is

$$\gamma = 1 - \rho_2 \approx \frac{2n}{n^2 + 1}. \tag{4.10}$$

The above specifies the maximum theoretically possible transmittance of an optical plate.

In the above example, transmittance of a glass plate is 88.6% (visible), while transmittance of a germanium plate is 41% (far IR). In the visible range, germanium transmittance is zero, which means that 100% of light is reflected and absorbed. Figure 4.5 shows reflectance and transmittance of a thin plate as functions of refractive indices. Here, a plate means any optical device (like a window or a lens) operating within its useful spectral range, that is, where its absorptive loss is small ($\alpha \approx 0$).

Figure 4.6 shows a light energy distribution within an optical plate when incident light flux Φ_0 strikes its surface. A part of incident flux Φ_ρ is reflected, another part Φ_α is absorbed by the material, and the third part Φ_γ is transmitted through. The absorbed portion of light is converted into heat, a portion of which ΔP is lost to

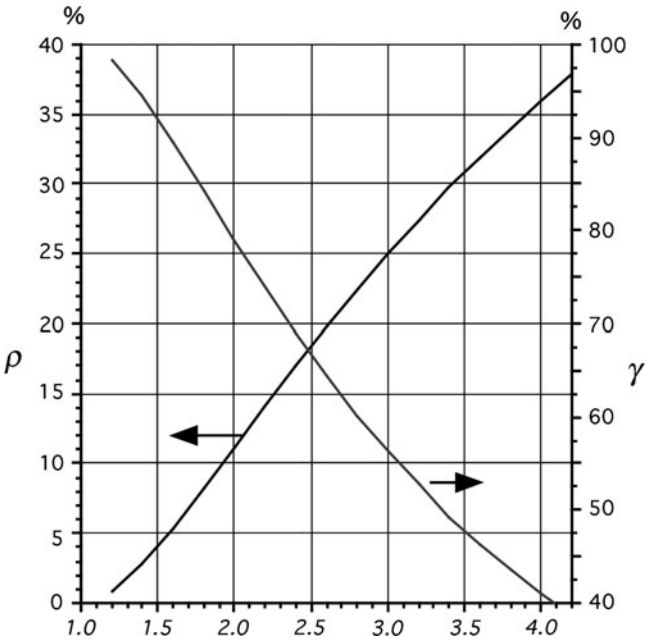

Fig. 4.5 Reflectance and transmittance of a thin plate as functions of a refractive index

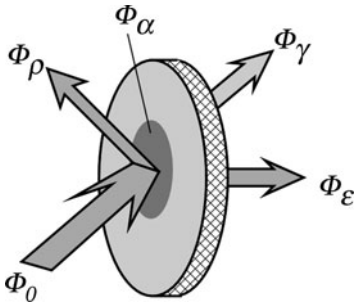

Fig. 4.6 Radiant energy distribution at optical plate

a supporting structure and surroundings through thermal conduction and convection. The rest of the absorbed light raises temperature of the material. The temperature increase may be of concern when the material is used as a window in a powerful laser. Another application where temperature increase may cause problems is in far-infrared detectors. The problem is associated with the flux $\Phi_\varepsilon = \Phi_\alpha - \Delta P$, which is radiated by the material due to its temperature change. This is called a secondary radiation. Naturally, a radiated spectrum relates to a temperature of the material and is situated in the far infrared region of the spectrum. The spectral

distribution of the secondary radiation corresponds to the absorption distribution of the material because absorptivity and emissivity are the same.

For materials with low absorption, the absorption coefficient can be determined through a temperature rise in the material:

$$\alpha = \frac{mc}{\Phi\gamma}\frac{2n}{n^2+1}\left(\frac{dT_g}{dt}+\frac{dT_L}{dt}\right)T_0, \tag{4.11}$$

where m and c are the mass and the specific heat of the optical material, T_g and T_L are the slopes of the rising and lowering parts of the temperature curve of the material, respectively, at test temperature T_0. Strictly speaking, light in the material is lost not only due to absorption but to scattering as well. A combined loss within material depends on its thickness and can be expressed through the so-called attenuation coefficient g and the thickness of the sample h. The transmission coefficient can be determined from (4.10), which is modified to account for the attenuation:

$$\gamma \approx (1-\rho_2)e^{-gh}. \tag{4.12}$$

The attenuation (or extinction) coefficient g is usually specified by manufacturers of optical materials.

4.2 Photometry

When using light-sensitive devices (photodetectors), it is critical to take into consideration both the sensor and light sources. In some applications, light is received from independent sources, while in others the light source is part of the measurements system. In any event, the so-called photometric characteristics of the optical system should be accounted for. Such characteristics include light, emittance, luminance, brightness, etc.

To measure radiant intensity and brightness, special units have been devised. Radiant flux (energy emitted per unit time), which is situated in a visible portion of the spectrum, is referred to as luminous flux. This distinction is due to the inability of the human eye to respond equally to like power levels of different visible wavelengths. For instance, one red and one blue light of the same intensity will produce very different sensations: the red will be perceived as much brighter. Hence, when comparing lights of different colors, the watt becomes a poor measure of brightness and a special unit called a lumen was introduced. It is based on a standard radiation source with molten platinum formed in a shape of a blackbody and visible through a specified aperture within a solid angle of one steradian. A solid angle is defined in a spherical geometry as

$$\omega = \frac{A}{r^2}, \tag{4.13}$$

where r is the spherical radius and A is the spherical surface of interest. When $A = r$, then the unit is called a spherical radian or *steradian* (see Table 1–8).

Illuminance is given as

$$E = \frac{\mathrm{d}F}{\mathrm{d}A}, \tag{4.14}$$

that is, a differential amount of luminous flux (F) over a differential area (A). It is most often expressed in lumens per square meter (square foot), or foot-meter (foot-candle). The luminous intensity specifies flux over solid angle:

$$I_L = \frac{dF}{d\omega}, \tag{4.15}$$

most often it is expressed in lumens per steradian or candela. If the luminous intensity is constant with respect to the angle of emission, the above equation becomes

$$I_L = \frac{F}{\omega}. \tag{4.16}$$

If the wavelength of the radiation varies, but the illumination is held constant, the radiative power in watts is found to vary. A relationship between illumination and radiative power must be specified at a particular frequency. The point of specification has been taken to be at a wavelength of 0.555 μm, which is the peak of the spectral response of a human eye. At this wavelength, 1 W of radiative power is equivalent to 680 lumens. For the convenience of the reader, some useful terminology is given in Table 4.1.

In the selection of electro-optical sensors, design considerations of light sources are of prime concern. A light source will effectively appear as either a point source or an area source, depending upon the relationship between the size of the source and the distance between the source and the detector. Point sources are arbitrarily defined as those whose diameter is less than 10% of the distance between

Table 4.1 Radiometric and photometric terminology

Description	Radiometric	Photometric
Total flux	Radiant flux (F) in watts	Luminous flux (F) in lumens
Emitted flux density at a source surface	Radiant emittance (W) in W/cm^2	Luminous emittance (L) in Lumens/cm^2 (Lamberts) or lumens/ft^2 (foot-lamberts)
Source intensity (point source)	Radiant intensity (I_r) in W/steradian	Luminous intensity (I_L) in Lumens/steradian (candela)
Source intensity (area source)	Radiance (B_r) in W/steradian/cm^2	Luminance (B_L) in Lumens/steradian/cm^2 (Lambert)
Flux density incident on a receiver surface	Irradiance (H) in W/cm^2	Illuminance (E) in lumens/cm^2 (candle) or lumens/ft^2 (foot-candle)

the source and the detector. While it is usually desirable that a photodetector is aligned such that its surface area is tangent to the sphere with the point source at its center, it is possible that the plane of the detector can be inclined from the tangent plane. Under this condition, the incident flux density (irradiance) is proportional to the cosine of the inclination angle φ:

$$H = \frac{I_r}{\cos\varphi}, \tag{4.17}$$

and the illuminance

$$E = \frac{I_L}{r^2}\cos\varphi. \tag{4.18}$$

The area sources are arbitrarily defined as those whose diameter is greater than 10% of the separation distance. A special case that deserves some consideration occurs when radius R of the light source is much larger than the distance r to the sensor. Under this condition

$$H = \frac{B_r A_s}{r^2 + R^2} \approx \frac{B_r A_s}{R^2}, \tag{4.19}$$

where A_s is the area of the light source and B_r is the radiance. Since the area of the source $A_s = \pi R^2$, irradiance is

$$H \approx B_\mathrm{r}\pi = W, \tag{4.20}$$

that is, the emitted and incident flux densities are equal. If the area of the detector is the same as area of the source, and $R \gg r$, the total incident energy is approximately the same as the total radiated energy, that is, unity coupling exists between the source and the detector. When the optical system is comprised of channeling, collimating, or focusing components, its efficiency and, subsequently, coupling coefficient must be considered. Important relationships for point and area light sources are given in Tables 4.2 and 4.3.

Table 4.2 Point source relationships

Description	Radiometric	Photometric
Point source intensity	I_r, W/sr	I_L lumens/sr
Incident flux density	Irradiance, $H = \frac{I_\mathrm{r}}{r^2}$, W/m^2	Illuminance, $E = \frac{I_\mathrm{L}}{r^2}$, lumens/m^2
Total flux output of a point source	$P = 4\pi I_\mathrm{r}$, W	$F = 4\pi I_\mathrm{L}$, lumens

Table 4.3 Area source relationships

Description	Radiometric	Photometric
Point source intensity	B_r, W/(cm^2·sr)	B_L lumens/(cm^2·sr)
Emitted flux density	$W = \pi B_\mathrm{r}$, W/cm^2	$L = \pi B_\mathrm{L}$, lumens/cm^2
Incident flux density	$H = \frac{B_r A_s}{r^2 + R^2}$, W/cm^2	$E = \frac{B_L A_s}{r^2 + R^2}$, lumens/cm^2

4.3 Windows

The main purpose of windows is to protect interiors of sensors and detectors from environment. A good window should transmit light rays in a specific wavelength range with minimal distortions. Therefore, windows should possess appropriate characteristics depending on a particular application. For instance, if an optical detector operates under water, perhaps its window should posses the following properties: a mechanical strength to withstand water pressure, low water absorption, a transmission band corresponding to the wavelength of interest, and an appropriate refractive index, which preferably should be close to that of water. A useful window that can withstand high pressures is spherical as shown in Fig. 4.7. To minimize optical distortions, three limitations should be applied to a spherical window: aperture D (its largest dimension) should be smaller than the window's spherical radius R_1, thickness d of the window should be uniform and much smaller than radius R_1. If these conditions are not met, the window becomes a concentric spherical lens.

Surface reflectivity of a window should be considered for its overall performance. To minimize a reflective loss, windows may be given antireflective coatings (ARC), which may be applied on either one or both sides of the window. These are the coatings that give bluish and amber appearances to photographic lenses and filters. Due to refraction in the window (see Fig. 4.4), a passing ray is shifted by a distance L, which for small angles Θ_1 may be found from formula:

$$L = d\frac{n-1}{n}, \tag{4.21}$$

where n is the refractive index of the material.

Sensors operating in the mid and far infrared ranges require special windows that are opaque in the visible and ultraviolet (UV) spectral regions and quite transparent in the wavelength of interest. Several materials are available for fabrication of such windows. Spectral transmittances of some materials are shown in Fig. 4.3 When

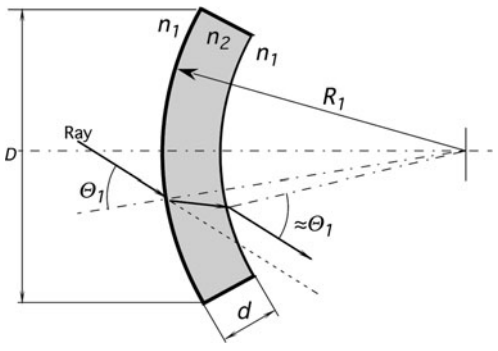

Fig. 4.7 Spherical window

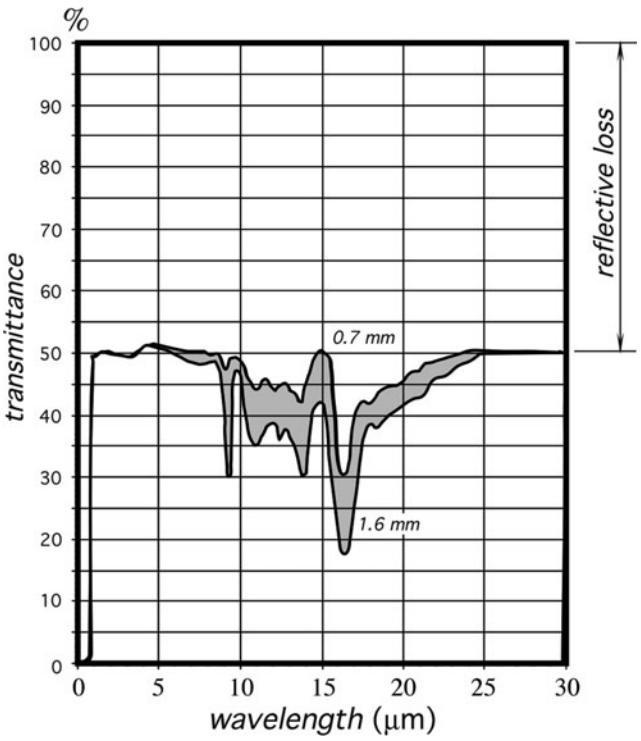

Fig. 4.8 Spectral transmittance of a silicon window. Note that majority of loss is due to a reflection from two surfaces

selecting material for a mid- and far-infrared window, the refractive index must be seriously considered because it determines the coefficient of reflectivity, absorptivity, and eventually transmittance. Figure 4.8 shows spectral transmittances of two silicon windows having different thicknesses. Total radiation (100%) at the window is divided into three portions: Reflected (about 50% over the entire spectral range), absorptive (varies at different wavelengths) and transmitted, which is whatever is left after the reflection and absorption. Since all windows are characterized by specific spectral transmissions, often they are called *filters*.

4.4 Mirrors

Mirror is the oldest optical instrument ever used or designed. Whenever light passes from one medium to another, there is some reflection. To enhance a reflectivity, a single or multilayer reflecting coating is applied on either the front (first surface) or the rear (second surface) of a plane-parallel plate or other substrate

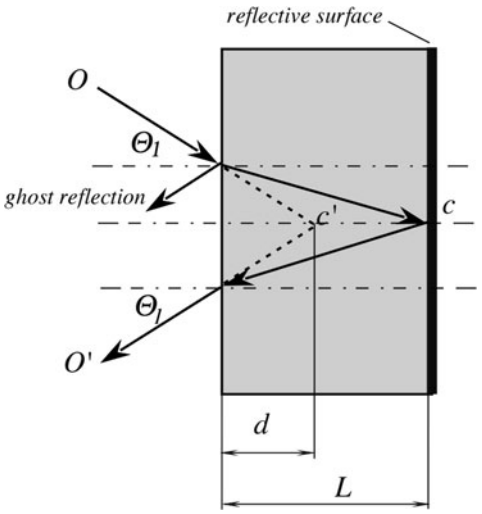

Fig. 4.9 Second surface mirror

of any desirable shape. The first surface mirrors are the most accurate. In the second surface mirror, light must enter a plate having generally a different index of refraction than the outside medium. A second surface mirror in effect is a combination of mirror and window.

Several effects in the second surface mirror must be taken into consideration. First, due to the refractive index n of a plate, a reflective surface appears closer (Fig. 4.9). A virtual thickness d of the carrier for smaller angles Θ_1 may be found from a simple formula:

$$d \approx \frac{L}{n}. \tag{4.22}$$

A front side of the second surface mirror may also reflect a substantial amount of light creating the so-called ghost reflection. A glass plate typically reflects about 4% of visible light. Further, a carrier material may have a substantial absorption in the wavelength of interest. For instance, if a mirror operates in a far infrared spectral range, it should use either first surface metallization or second surface where the substrate is fabricated of ZnSe or other long wavelength transparent materials. Materials such as Si or Ge have too strong surface reflectivity to be useful for the fabrication of the second surface mirrors.

Reflecting coatings applied to a surface for operation in the visible and near infrared range can be silver, aluminum, chromium, and rhodium. Gold is preferable for the mid- and far-infrared spectral range devices. By selecting an appropriate coating, the reflectance may be achieved of any desired value from nearly 0 to almost 1 (Fig. 4.10).

The best mirrors for broadband use have pure metallic layers, vacuum or electrolytically deposited on glass, fused silica, or metal substrates. Before the

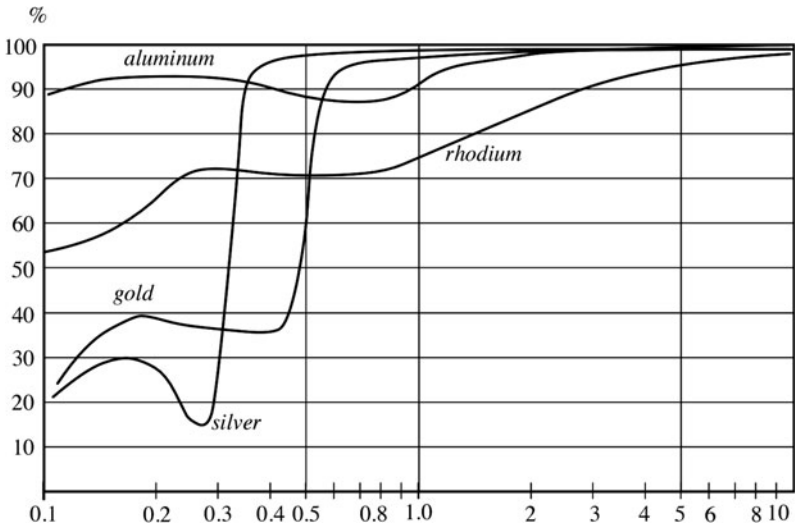

Fig. 4.10 Spectral reflectances of some mirror coatings

reflective layer deposition, to achieve a leveling effect, a mirror may be given an undercoat of copper, zirconium-copper or molybdenum.

Another useful reflector, which may serve as a second surface mirror without the need for reflective coatings, is a prism where the effect of total internal reflection (TIR) is used. The angle of a total internal reflection is a function of a refractive index:

$$\Theta_0 = \arcsin\left(\frac{1}{n}\right) \tag{4.23}$$

The total internal reflectors are the most efficient in the visible and near infrared spectral ranges as the reflectivity coefficient is close to unity. The TIR principle is fundamental for the operation of optical fibers.

A reflective surface may be formed practically in any shape to divert the direction of light travel. In the optical systems, curved mirrors produce effects equivalent to that of lenses. The advantages they offer include (1) higher transmission, especially in the longer wavelength spectral range where lenses become less efficient due to higher absorption and reflectance loss, (2) absence of distortions incurred by refracting surfaces due to dispersion (chromatic aberrations), and (3) lower size and weight when compared with many types of lenses. Spherical mirrors are used whenever light must be collected and focused.[1] However, spherical

[1] *Focus* is from the Latin meaning *fireplace*, a gathering place in a house.

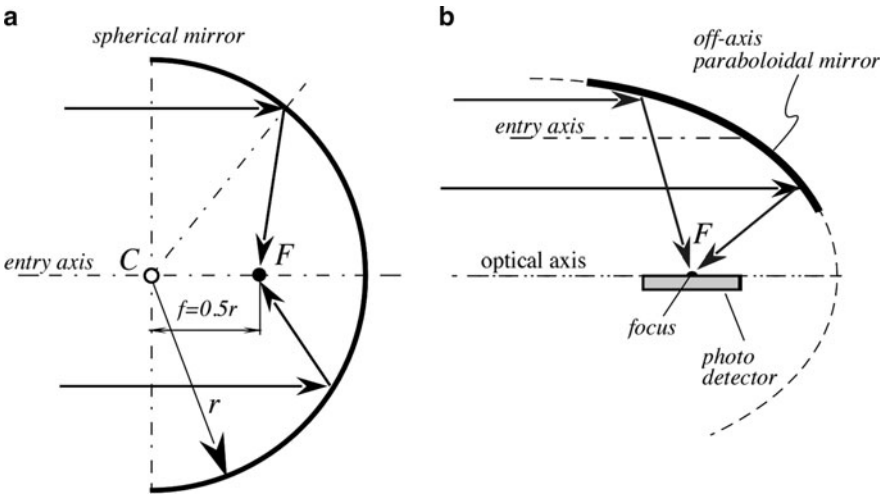

Fig. 4.11 Spherical (**a**) and parabolic (**b**) first surface mirrors

mirrors are good only for the parallel or near parallel beams of light that strike a mirror close to normal. These mirrors suffer from imaging defects called aberrations. Figure 4.11a shows a spherical mirror with the center of curvature in point *C*. A focal point is located at a distance of 1/2 of the radius from the mirror surface. A spherical mirror is astigmatic, which means that the off-axis rays are focused away from its focal point. Nevertheless, such mirrors prove very useful in detectors where no quality imaging is required, for instance in infrared motion detectors which are covered in detail in Sect. 6.5.

A parabolic mirror is quite useful for focusing light off-axis. When it is used in this way, there is complete access to the focal region without shadowing, as shown in Fig. 4.11b.

4.5 Lenses

Lenses[2] are useful in sensors and detectors to divert the direction of light rays and arrange them in a desirable fashion. Figure 4.12 shows a planoconvex lens, which has one spherical surface and the other is flat. The lens has two focuses at both sides: F and F', which are positioned at equal distances $-f$ and f from the lens.

[2]The word *lens* is from the Latin name for lentils. A lentil seed is flat and round, as its sides bulge outward – just like a convex lens.

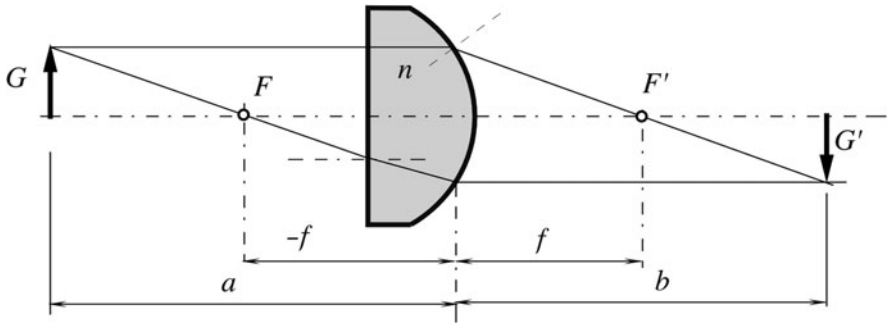

Fig. 4.12 Geometry of a plano-convex lens

When light rays from object G enters the lens, their directions change according to Snell's law.

To determine the size and the position of an image created by the lens, it is convenient to draw two rays that have special properties. One is parallel to the optical axis, which is a line passing through the sphere's center of curvature. After exiting the lens that ray goes through focus F'. The other ray first goes through focus F and upon exiting the lens, propagates in parallel with the optical axis. A thin lens whose radius of curvature is much larger than the thickness of lens has a focal distance f that may be found from the equation

$$\frac{1}{f} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right), \tag{4.24}$$

where r_1 and r_2 are radii of the lens curvatures. Image G' is inverted and positioned at a distance b from the lens. That distance may be found from a thin lens equation

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b}. \tag{4.25}$$

For the thick lenses where thickness t is comparable with the radii of curvature, a focal distance may be found from formula

$$f = \frac{nr_1r_2}{(n-1)|n(r_1+r_2) - t(n-1)|}. \tag{4.26}$$

Several lenses may be combined into a more complex system. For two lenses separated by distance d, combination focal length may be found from equation

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}. \tag{4.27}$$

4.6 Fresnel Lenses

Fresnel lenses are optical elements with step-profiled surfaces. They prove to be very useful in sensors and detectors where a high quality of focusing is not required and primarily the light energy is of prime concern. Major applications include light condensers, magnifiers, and focusing element in occupancy detectors. Fresnel lenses may be fabricated of glass, acrylic (visible and near infrared range), or polyethylene (mid- and far-infrared ranges). The history of Fresnel lenses began in 1748, when Count Buffon proposed grinding out a solid piece of glass lens in steps of the concentric zones in order to reduce the thickness of the lens to a minimum and to lower energy loss. He realized that only the surface of a lens is needed to refract light, because once the light is inside the lens, it travels in a straight line. His idea was modified in 1822 by Augustin Fresnel (1788–1827), who constructed a lens in which the centers of curvature of the different rings receded from the axis according to their distances from the center, so as to practically eliminate spherical aberration.

The concept of that lens is illustrated in Fig. 4.13, where a regular planoconvex lens is depicted. The lens is sliced into several concentric rings. After slicing, all rings still remain lenses, which refract incident rays into a common focus defined by (4.24). A change in an angle occurs when a ray exits a curved surface, not inside the lens; hence, the section of a ring marked by the letter x does not contribute to the focusing properties. If all such sections are removed, the lens will look like as it is shown in Fig. 4.13b and will fully retain its ability to focus light rays. Now, all of the rings may be shifted with respect to one another to align their flat surfaces (Fig. 4.13c). A resulting near-flat but grooved lens is called Fresnel, which has nearly the same focusing properties as the original planoconvex lens.

A Fresnel lens basically consists of a series of concentric prismatic grooves designed to cooperatively direct incident light rays into a common focus. It has several advantages over conventional lens, such as low weight, thin size, ability to be curved (for a plastic lens) to any desirable shape and, most importantly, a lower absorption loss of the light flux. This is the prime reason why this type of a lens is almost has exclusively been used in lighthouses to form parallel beams of light (Fig. 4.14). A lower absorption loss is very important for fabrication of the mid and

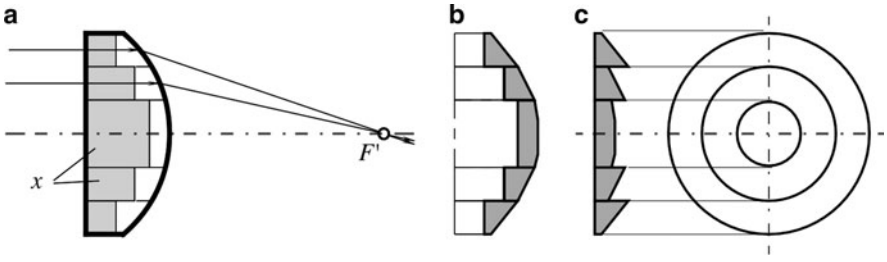

Fig. 4.13 Concept of a Fresnel lens.

Fig. 4.14 Fresnel lens for a lighthouse

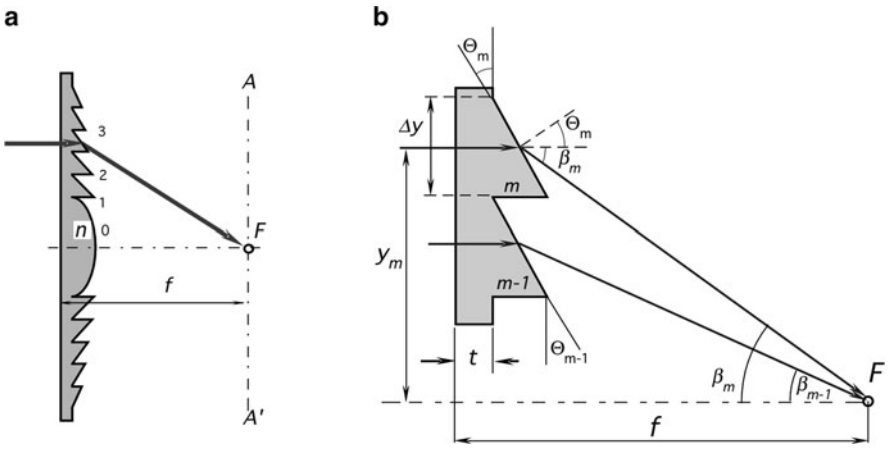

Fig. 4.15 Grooves of the Fresnel lens (**a**); computation of the groove angle (**b**)

far infrared lenses where absorption in the material may be significant. This is the reason why low-cost polymer Fresnel is used almost exclusively in the far-infrared motion detectors.

When fabricating a Fresnel lens, it is difficult to maintain a curved surface of each small groove; hence, the profile of a groove is approximated by a flat surface (Fig. 4.15a). This demands that the steps be positioned close to each other. In fact, the closer the steps, the more accurate the lens. The limiting factor is the ability to tool and fabricate such closely positioned grooves. There are several ways of designing grooves of the lens. The most common is the so-called constant step where all groves have the same pitch, that is, the distance between the neighboring grooves.

A computation of the lens is essentially the computation of a groove angle, depending on its number [14]. We assume that a monochromatic parallel beam is

incident normally from the left onto a flat surface of the lens. The refraction takes place only at the grooved side. Applying Snell's law of refraction of a ray passing through the center of the groove, we arrive at

$$\sin\Theta_m = n\sin\left(\Theta_m + \beta_m\right), \tag{4.28}$$

where n is the refractive index of the material of the desired wavelength, and the angles are defined in Fig. 4.15b. Let us consider y_m be a distance from the optical axis to the m-th groove, then for that particular groove

$$\Theta_m = \tan^{-1}\left[\frac{y_m}{n\sqrt{y_m^2 + (f-t)^2} - (f-t)}\right], \tag{4.29}$$

where f is the focal length and t is the mean lens thickness. This equation can be rewritten in a dimensionless form, considering

$$y'_m = \frac{y_m}{f} \quad \text{and} \quad t' = \frac{t}{f} \tag{4.30}$$

and finally we arrive at a basic formula for computing a Fresnel lens:

$$\Theta_m = \tan^{-1}\left[\frac{y'_m}{n\sqrt{y'^2_m + (1-t')^2} - (1-t')}\right]. \tag{4.31}$$

The angles Θ'_m of the refracting prisms are fixed such that all the central rays of a particular wavelength have a common focus. A refractive index n may be found from Table A.19. For the mid- and far-infrared ranges, low-density polyethylene (LDPE) has refractive index 1.510, while the high-density polyethylene (HDP) has $n = 1.540$.

A Fresnel lens may be slightly bent if it is required for a sensor design. However, a bend changes the positions of focal points. If a lens is bent with its groves inside the curvature, all angles Θ'_m change depending on the radius of curvature. A new focal distance can be found from inverting (4.29) and solving it for f.

4.7 Fiber Optics and Waveguides

Although light does not go around the corner, it can be channeled along complex paths by the use of waveguides. To operate in the visible and near infrared spectral ranges, the guides may be fabricated of glass or polymer fibers. For the mid- and

far-infrared spectral ranges, the waveguides are made of special materials or as hollow tubes with highly reflective inner surfaces. The tubular waveguide operates on the principle of reflection where light beams travel in a zigzag pattern. A fiber can be used to transmit light energy in the otherwise inaccessible areas without any transport of heat from the light source.

The surface and ends of a round or other cross-section fiber are polished. An outside cladding may be added. When glass is hot, the fibers can be bent to curvature radii of 20–50 times their section diameter and after cooling, to 200–300 diameters. Plastic fibers fabricated of polymethyl methacrylate may be bent at much smaller radii than glass fibers. A typical attenuation for a 0.25-mm polymer fiber is in the range of 0.5 dB/m of length. Light propagates through a fiber by means of a total internal reflection, as shown in Fig. 4.16b. It follows from (4.23) that light passing to air from a medium having a refractive index n is subject to the limitation of an angle of total internal reflection. In a more general form, light may pass to another medium (cladding) having refractive index n_1, then, (4.23) becomes

$$\Theta_0 = \arcsin\left(\frac{n_1}{n}\right) \tag{4.33}$$

Figure 4.16a shows a profile of the index of refraction for a single fiber with the cladding where the cladding must have a lower index of refraction to assure a total internal reflection at the boundary. For example, a silica-clad fiber may have compositions set so that the core (fiber) material has an index of refraction of 1.5, and the clad has an index of refraction of 1.485. To protect the clad fiber, it is typically enclosed in some kind of protective rubber or plastic jacket. This type of the fiber is called a "step index multimode" fiber, which refers to the profile of the index of refraction.

When light enters the fiber, it is important to determine the maximum angle of entry which will result in total internal reflections (Fig. 4.16b). If we take that

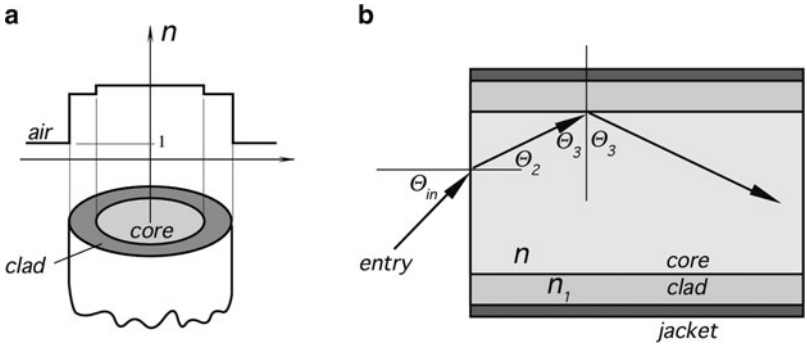

Fig. 4.16 Optical fibers: A step-index multiple fiber (**a**) and determination of the maximum angle of entry (**b**)

minimum angle of an internal reflection $\Theta_0 = \Theta_3$, then the maximum angle Θ_2 can be found from Snell's law:

$$\Theta_{2(\max)} = \arcsin\left(\frac{\sqrt{n^2 - n_1^2}}{n}\right), \tag{4.34}$$

Applying Snell's law again and remembering that for air $n \approx 1$, we arrive at

$$\sin\Theta_{\text{in}(\max)} = n_1 \sin\Theta_{2(\max)} \tag{4.35}$$

Combining (4.34) and (4.35), we obtain the largest angle with the normal to the fiber end for which the total internal reflection will occur in the core:

$$\Theta_{in(\max)} = \arcsin\left(\sqrt{n^2 - n_1^2}\right). \tag{4.36}$$

Light rays entering the fiber at angles greater than $\Theta_{\text{in}(\max)}$ will pass through to the jacket and will be lost. For data transmission, this is an undesirable event. However, in a specially designed fiber-optic sensor, the maximum entry angle can be a useful phenomenon for modulating light intensity. Sometimes, the value $\Theta_{\text{in}(\max)}$ is called the numerical aperture of the fiber. Due to variations in the fiber properties, bends, and skewed paths, the light intensity does not drop to zero abruptly but rather gradually diminishes to zero while approaching $\Theta_{\text{in}(\max)}$. In practice, the numerical aperture is defined as the angle at which light intensity drops by some arbitrary number (e.g. −10 dB of the maximum value).

One of the useful properties of fiber-optic sensors is that they can be formed into a variety of geometrical shapes depending on the desired application. They are very useful for the design of miniature optical sensors which are responsive to such stimuli, as pressure, temperature, chemical concentration, and so forth. The basic idea for the use of fiber optics in sensing is to modulate one or several characteristics of light in a fiber and, subsequently, to optically demodulate the information by conventional methods.

A stimulus may act on a fiber either directly or it can be applied to a component attached to the fiber's outer surface or the polished end to produce an optically detectable signal. To make a fiber chemical sensor, a special solid phase of a reagent may be formed in the optical path coupled to the fiber. The reagent interacts with the analyte to produce an optically detectable effect (e.g., modulating the index of refraction or coefficient of absorption). A cladding on a fiber may be created from a chemical substance whose refractive index may be changed in the presence of some fluids [3]. When the angle of total internal reflection changes, the light intensity varies.

Optical fibers may be used in two modes. In the first mode (Fig. 4.17a), the same fiber is used to transmit the excitation signal and to collect and conduct an optical response back to the processing device. In the second mode, two or more fibers

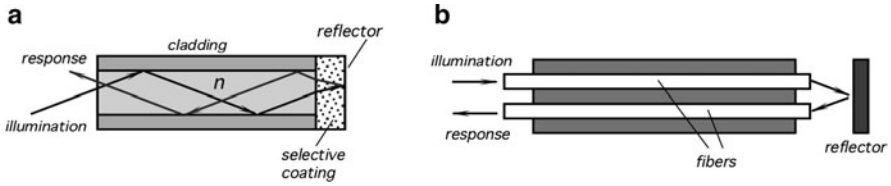

Fig. 4.17 Single (**a**) and dual (**b**) fiber-optic sensors.

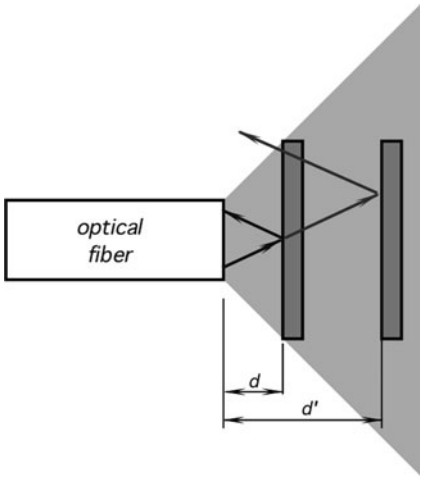

Fig. 4.18 Fiber-optic displacement sensor utilizes the modulation of reflected light intensity

are employed where excitation (illumination) function and collection function are carried out by separate fibers (Fig. 4.17b). The most commonly used type of fiber-optic sensor is an intensity sensor, where light intensity is modulated by an external stimulus [4]. Figure 4.18 shows a displacement sensor where a single-fiber waveguide emits light toward the reflective surface. Light travels along the fiber and exits in a conical profile toward the reflector. If the reflector is close to the fiber end (distance d), most of the light is reflected into the fiber and propagates back to the light detector at the other end of the fiber. If the reflector moves away, some of the rays are reflected outside of the fiber end, and fewer photons are returned back. Due to a conical profile of the emitted light, a quasilinear relationship between the distance d and the intensity of the returned light can be achieved over a limited range.

The so-called microbend strain gauge can be designed with an optical fiber, which is squeezed between two deformers, as shown in Fig. 4.19a. The external force applied to the upper deformer bends the fiber, affecting a position of an internal reflective surface. Thus, a light beam, which normally would be reflected in direction x, approaches the lower part of the fiber at an angle which is less than Θ_0, the angle of total internal reflection (4.33). Thus, instead of being reflected, light is refracted and moves in the direction y through the fiber wall. The closer the deformers come to each other, the more light goes astray and the less light is transmitted along the fiber.

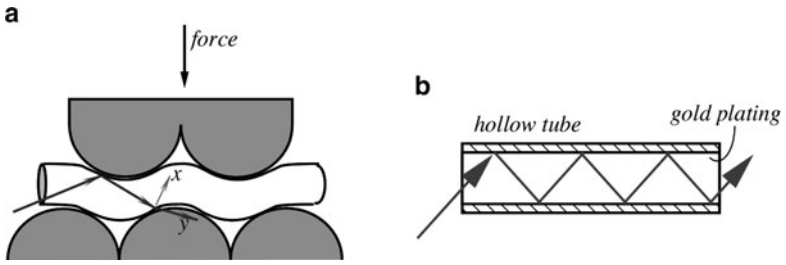

Fig. 4.19 Fiber-optic microbend strain gauge (**a**) and a waveguide for the far infrared radiation (**b**)

For operation in the spectral range where loss in fibers is too great (mid- and far-infrared spectral ranges), hollow tubes are generally used for light channeling (Fig. 4.19b). The tubes are highly polished inside and coated with reflective metals. For instance, to channel thermal radiation, a tube may be fabricated of brass and coated inside by two layers: Nickel as a leveling underlayer and the optical quality gold having thickness in the range 500–1,000 Å. Hollow waveguides may be bent to radii of 20 or more of their diameters.

Although fiber optics use the effect of the total internal reflection, tubular waveguides use a first surface mirror reflection, which is always less than 100%. As a result, loss in a hollow waveguide is a function of a number of reflections; that is, loss is higher for the smaller diameter and the longer length of a tube. At length/diameter ratios more than 20, hollow waveguides become quite inefficient and fiber optic devices should be considered; for example, AMTIR (Table A.26).

4.8 Concentrators

There is an important issue of increasing density of the photon flux impinging on the sensor's surface. In many cases, when only the energy factors are of importance, and a focusing or imaging is not required, special optical devices can be used quite effectively. These are the so-called nonimaging collectors, or concentrators [5, 16]. They have some properties of the waveguides and some properties of the imaging optics (like lenses and curved mirrors). The most important characteristic of a concentrator is the ratio of the area of the input aperture divided by the area of the output aperture, the concentration ratio C. Its value is always more than unity. That is, the concentrator collects light from a larger area and directs it to a smaller area [Fig. 4.20a] where the sensing element is positioned. There is a theoretical maximum for C:

$$C_{\max} = \frac{1}{\sin^2 \Theta_i}, \tag{4.37}$$

where Θ_i is the maximum input semiangle. Under these conditions, the light rays emerge at all angles up to $\pi/2$ from the normal to the exit face. This means

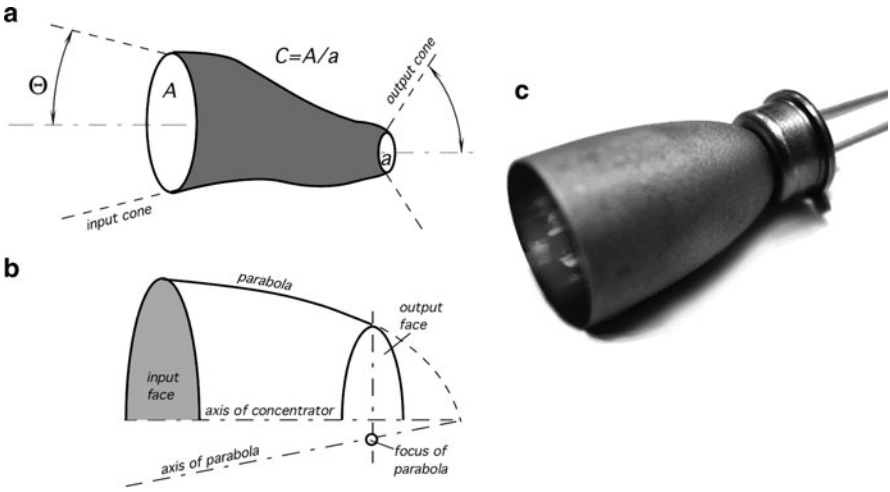

Fig. 4.20 Nonimaging concentrator: General schematic (**a**), concentrator having a parabolic profile (**b**) and Winston cone attached to a pyroelectric sensor

that the exit aperture diameter is smaller by $\sin\Theta_i$ times the input aperture. This gives an advantage in the sensor design as its linear dimensions can be reduced by that number while maintaining a near equal efficiency. The input rays entering at angle Θ will emerge within the output cone with angles dependent of point of entry.

The concentrators can be fabricated with reflective surfaces (mirrors) or refractive bodies (e.g., Fresnel lenses), or as combinations of both. A practical shape of the reflective parabolic concentrator is shown in Fig. 4.20b and its connection to a sensor in Fig. 4.20c. It is interesting to note that cone light receptors in the retina of a human eye have a shape similar to that shown in Fig. 4.20b [6].

The tilted parabolic concentrators may have very high efficiency; they can collect and concentrate well over 90% of the incoming radiation.[3] If a lesser efficiency is acceptable, a conical rather than paraboloid concentrator can be employed. Some of the incoming rays will be turned back after several reflections inside the cone, however, its overall efficiency is still near 80%. Clearly, the cones are easier to fabricate than the paraboloids of revolution.

4.9 Coatings for Thermal Absorption

All thermal radiation sensors rely on absorption or emission of the electromagnetic waves in the mid- and far-infrared spectral ranges. According to Kirchhoff's discovery, absorptivity α and emissivity ε is the same thing (see Sect. 3.12.3 of Chap. 3).

[3]These concentrators sometimes are called Winston cones.

Their value for the efficient sensor's operation must be maximized, i.e., it should be made as close to unity as possible. This can be achieved by either processing the surface of a sensor to make it highly emissive or covering it with a special coating having a high emissivity. Any such coating should have a good thermal conductivity and a very small thermal capacity, which means that it must be thin.

Several methods are known to give a surface the emissive (absorptive) properties. Some of them are deposition of thin metal films (like nichrome) having reasonably good emissivity, galvanic deposition of porous platinum black [7] and evaporation of metal in atmosphere of low-pressure nitrogen [8]. The most effective way to create a highly absorptive (emissive) material is to form it with a porous surface [9]. Particles with sizes much smaller than the wavelength generally absorb and diffract light. High emissivity of a porous surface covers a broad spectral range; however, it decreases with the increased wavelength. A film of goldblack with a thickness corresponding to 500 μg/cm^2 has an emissivity over 0.99 in the near, mid-, and far-infrared spectral ranges.

To form porous platinum black, the following electroplating recipe can be used [10]:

Platinum chloride	H_2PTCl_6 aq	2 g
Lead acetate	$Pb(OOCCH_3)_2 \cdot 3H_2O$	16 mg
Water	H_2O	58 g

Out of this galvanic bath, the films were grown at room temperature on silicon wafers with a gold underlayer film. A current density was 30 mA/cm^2. To achieve absorption better than 0.95, a film of 1.5 g/cm^2 is needed.

To form a goldblack by evaporation, the process is conducted in a thermal evaporation reactor in a nitrogen atmosphere of 100 Pa pressure. The gas is injected via a microvalve, and the gold source is evaporated from the electrically heated tungsten wire from a distance of about 6 cm. Due to collisions of evaporated gold with nitrogen, the gold atoms lose their kinetic energy and are slowed down to thermal speed. When they reach the surface, their energy is too low to allow surface mobility, and they stick to the surface on the first touch event. Gold atoms form a surface structure in the form of needles with linear dimensions of about 25 nm. The structure resembles a surgical cotton wool. For the best results, goldblack should have thickness in the range from 250 to 500 μg/cm^2.

Another popular method to enhance emissivity is to oxidize a surface metal film to form metal oxide, which generally is highly emissive. This can be done by a metal deposition in a partial vacuum.

Another method of improving the surface emissivity is to coat a surface with an organic paint (visible color of the paint is not important). These paints have far-infrared emissivity from 0.92 to 0.97; however, the organic materials have low thermal conductivity and cannot be effectively deposited with thicknesses less than 10 μm. This may significantly slow the sensor's speed response. In micromachined sensors, the top surface may be given a passivation glass layer, which not only

provides an environmental protection, but has emissivity of about 0.95 in the far infrared spectral range.

4.10 Nano-optics

Nano-optics deals with the interaction of light with particles or substances, at deeply subwavelength scales. Nano-structure-based optics, or nano-optics, is a class of optical devices based on finely patterned materials with critical dimensions several times smaller than the wavelength of light at which they are applied. By combining material and structural properties, nano-optics allow optical devices that are thin, offer high performance, and are highly reliable. Because nano-optic devices are produced using semiconductor-like manufacturing methods, they are readily integrated in arrays or multilayer structures, or with other optical and electronic components. This enables optical circuit designers to simplify optical circuits by combining optical functions or to increase functional capability by using tunable devices.

References

1. Begunov BN, Zakaznov NP, Kiryushin SI, Kuzichev VI (1988) *Optical instrumentation. theory and design*. Mir Publishers, Moscow
2. Applications of phototransistors in electro-optic systems. AN-508. © Motorola, 1988
3. Giuliani JF (1989) Optical waveguide chemical sensors. In: *Chemical sensors and micro-instrumentation*. Chapter 24, American Chemical Society, Washington
4. Mitchell GL (1991) Intensity-based and Fabry-Perot interferometer sensors. In: Udd E (ed) *Fiber optic sensors: an introduction for engineers and scientists*. Chapter 6. Wiley, New York
5. Welford WT, Winston R (1989) *High collection nonimaging optics*. Academic Press, San Diego, CA
6. Winston R, Enoch JM (1971) Retinal cone receptor as an ideal light collector. *J Opt Soc Am* 61:1120–21
7. von Hevisy G, Somiya T (1934) Über platinschwarz. *Zeitschrift für phys. Chemie A* 171:41
8. Harris L, McGinnes R, Siegel B (1948) *J Opt Soc Am* 38:7
9. Persky MJ (1999) Review of black surfaces for space-borne infrared systems. *Rev Sci Instrum* 70(5):2193–2217
10. Lang W, Kühl K, Sandmaier H (1991) Absorption layers for thermal infrared detectors. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators*. Digest of technical papers, pp. 635–638, IEEE
11. Yariv A (1985) *Optical electronics*, 3rd ed., Holt, Reinhart and Winston, New York
12. Johnson LM (1991) Optical modulators for fiber optic sensors. In: Udd E (ed) *Fiber optic sensors: introduction for engineers and scientists*. Wiley, New York
13. Haus HA (1984) *Waves and fields in optoelectronics*. Prentice-Hall, Englewood Cliffs, NJ
14. Sirohi RS (1979) Design and performance of plano-cylindrical Fresnel lens. *Appl Opt* 14, 45A (4):1509–1512
15. Katz M (1994) *Introduction to geometrical optics*. Penumbra Publishing Co.
16. Leutz R, Suzuki A (2001) *Nonimaging Fresnel lenses: design and performance of solar concentrators*, Springer, Berlin
17. Kingslake R (1978) *Lens design fundamentals*. Academic Press, New York

Chapter 5
Interface Electronic Circuits

Engineers like to solve problems.
If there are no problems handily available,
they will create their own problems.

-Scott Adams

5.1 Input Characteristics of Interface Circuits

A system designer is rarely able to connect a sensor directly to processing, monitoring, or recording instruments, unless a sensor has a built-in electronic circuit with an appropriate output format. When a sensor generates an electric signal, that signal often is either too weak, or too noisy, or it contains undesirable components. Besides, the sensor output may be not compatible with the input requirements of a data acquisition system, that is, it may have a wrong output format. To mate a sensor and a processing device, they either must share a "common value" or some kind of a "mating" device is required in-between. In other words, signal from a sensor usually has to be *conditioned* before it is fed into a processing device (a load). Such a load usually requires either voltage or current as its input signal.

An interface or a *signal conditioning* circuit has a specific purpose: to bring signal from the sensor up to the format that is compatible with the load device. Figure 5.1 shows a stimulus that acts on a sensor, which is connected to a load through an interface circuit. To do its job effectively, an interface circuit must be a faithful slave of two masters: the sensor and the load device. Its input characteristics must be matched to the output characteristics of the sensor and its output must be interfaceable with the load. This book, however, focuses on the sensors, therefore, below we will discuss only the front stages of the interface circuits. Also, we will discuss some typical excitation circuits that are required for active sensors, that is, for the sensors which need electrical signals to produce electrical outputs.

The input part of an interface circuit may be specified through several standard numbers. These numbers are useful for calculating how accurately the circuit can

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_5, © Springer Science+Business Media, LLC 2010

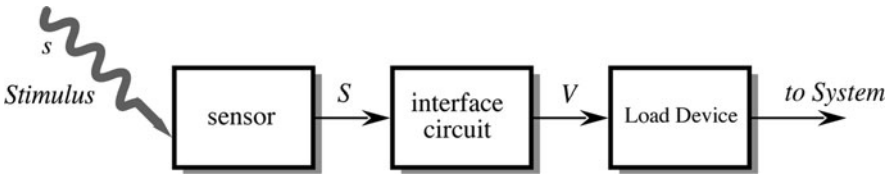

Fig. 5.1 Interface circuit matches the signal formats of a sensor and a load device

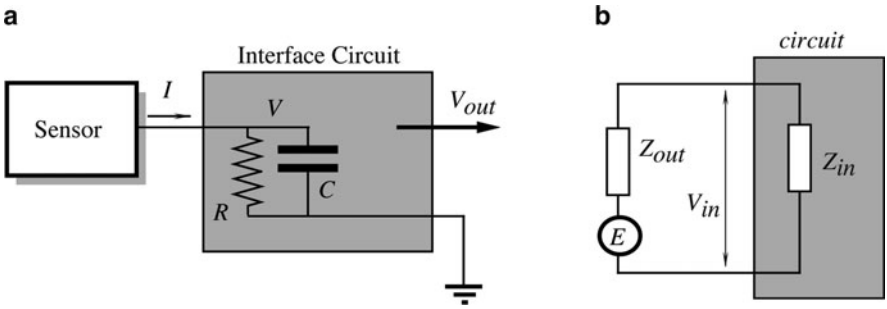

Fig. 5.2 Complex input impedance of an interface circuit (**a**), and equivalent circuit of a voltage generating sensor (**b**)

process the sensor's output signal and what would be the circuit's contribution to a total error budget?

The input impedance; shows by how much the circuit loads the sensor. The impedance may be expressed in a complex form as:

$$\mathbf{Z} = \frac{\mathbf{V}}{\mathbf{I}}, \tag{5.1}$$

where **V** and **I** are complex notations for the voltage and the current across the input impedance. For example, if the input of a circuit is modeled as a parallel connection of input resistance, R and input capacitance, C (Fig. 5.2a), the complex input impedance may be represented as

$$\mathbf{Z} = \frac{R}{1 + j\omega RC}, \tag{5.2}$$

where ω is the circular frequency and $j = \sqrt{-1}$ is the imaginary unity. At very low frequencies, a circuit having a relatively low input capacitance and resistance has an input impedance, which is almost equal to the input resistance: $\mathbf{Z} \approx R$. Relatively low, here it means that the reactive part of the above equation becomes small, i.e., the following holds:

$$RC << \frac{1}{\omega}. \tag{5.3}$$

Whenever an input impedance of a circuit is considered, the output impedance of the sensor must be taken into account. For example, if the sensor is of a capacitive nature, to define a frequency response of the input stage, sensor's capacitance must be connected in parallel with the circuit's input capacitance. Formula (5.2) suggests that the input impedance is function of the signal frequency. With an increase in the signal rate of change, the input impedance becomes lower.

Figure 5.2b shows an equivalent circuit for a voltage generating sensor. The circuit is comprised of the sensor output, $\mathbf{Z}_{out}$, and the circuit input, $\mathbf{Z}_{in}$, impedances. The output signal from the sensor is represented by a voltage source, e, which is connected in series with the output impedance. Instead of a voltage source, for some sensors it is more convenient to represent the output signal as outgoing from a current source, which would be connected in parallel with the sensor output impedance. Both representations are equivalent to one another, so we will use voltage. Accounting for both impedances, the circuit input voltage, V_{in} is represented as

$$V_{in} = e\frac{Z_{in}}{Z_{in} + Z_{out}}. \tag{5.4}$$

In any particular case, an equivalent circuit of a sensor should be defined. This helps to analyze the frequency response and the phase lag of the sensor-interface combination. For instance, a capacitive detector may be modeled as a pure capacitance connected in parallel with the input impedance. Another example is a piezoelectric sensor which can be represented by a very high resistance (on the order of 10^{11} Ω) shunted by a capacitance (in the order of 10pF).

To illustrate the importance of the input impedance characteristics, let us consider a purely resistive sensor connected to the input impedance as shown in Fig. 5.2. The circuit's input voltage as function of frequency, f, can be expressed by a formula

$$V = \frac{E}{\sqrt{1 + \left(\frac{f}{f_c}\right)^2}} \tag{5.5}$$

where $f_c = (2\pi RC)^{-1}$ is the corner frequency, (i.e., the frequency where the amplitude drops by 3 dB). If we assume that a 1% accuracy in the amplitude detection is required, then we can calculate the maximum stimulus frequency that can be processed by the circuit:

$$f_{max} \approx 0.14 f_c, \tag{5.6}$$

or $f_c \approx 7 f_{max}$; that is, the impedance must be selected in such a way as to assure a sufficiently high corner frequency. For example, if the stimulus' highest frequency is 100 Hz, the corner frequency must be selected at least at 700 Hz. In practice, f_c is selected even higher, because of the additional frequency limitations in the subsequent circuits.

One should not overlook a speed response of the front stage of the interface circuit. Operational amplifiers, which are the most often used building blocks of interface circuits, usually have limited frequency bandwidths. There are the so-called programmable operational amplifiers, which allow the user to control (to program) the bias current and, therefore, the first stage frequency response. The higher the current, the faster would be the response.

Figure 5.3 is a more detailed equivalent circuit of the input properties of an interface circuit, for instance, an amplifier or an A/D converter. The circuit is characterized by the input impedance $\mathbf{Z}_{in}$ and several generators. They represent voltages and currents that are generated by the circuit itself. These signals are spurious and may pose substantial problems if not handled properly. All these interfering signals are temperature-dependent.

Voltage e_o is called the input *offset voltage*. If the input terminals of the circuit are shorted together, that voltage would simulate a presence of an input dc signal having a value of e_o. It should be noted that the offset voltage source is connected in series with the input and its resulting error is independent of the output impedance of the sensor.

The input *bias current* i_o is also internally generated by the circuit. Its value is quite high for many bipolar transistors, much smaller for the JFET s, and even more lower for the CMOS circuits. This current may present a serious problem when a circuit or sensor employs high-impedance components. The bias current passes through the input impedance of the circuit and the output impedance of the sensor, resulting in a spurious voltage drop. This voltage may be of a significant magnitude. For instance, if a piezoelectric sensor is connected to a circuit having an input resistance of 1 GΩ (10^9 Ω) and the input bias current of 1 nA (10^{-9}A), the voltage drop at the input becomes equal to 1 GΩ·1 nA=1 V, a very high value indeed. In contrast to the offset voltage, the error resulting from bias current is proportional to the output impedance of the sensor. This error is negligibly small for the sensors having low output resistances. For instance, an inductive detector is not sensitive to a magnitude or variations in the bias current.

A circuit board *leakage current* may be a source of errors while working with high-impedance circuits. This current may be the result of lower surface resistance in the printed circuit board (PCB). Possible causes for that are: poor quality PCB material, surface contamination with solder flux residue (a poorly washed board),

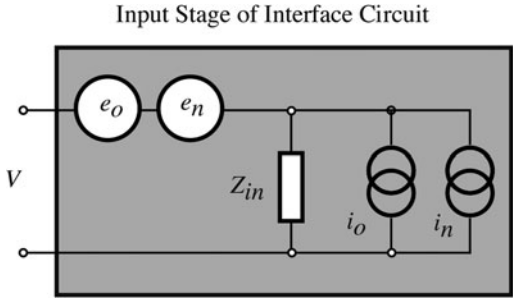

Fig. 5.3 Equivalent circuit of electrical noise sources at an input stage

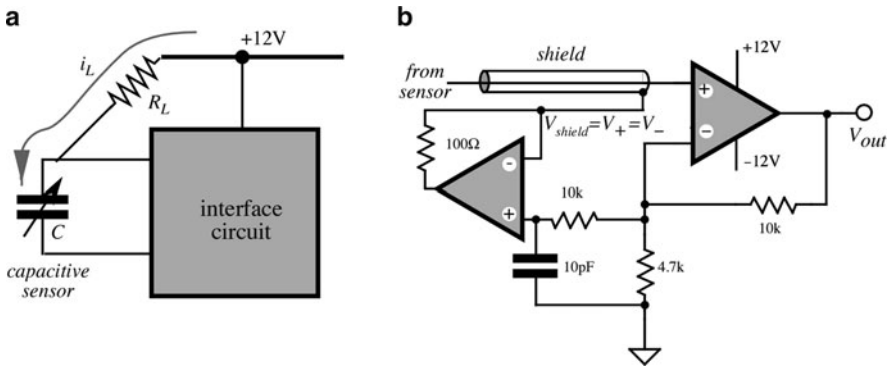

Fig. 5.4 Circuit board leakage affects input stage (**a**); driven shield of the input stage (**b**)

moisture, and degraded conformal coating. Figure 5.4a shows that a power supply bus and the board resistance, R_L, may cause leakage current, i_L, through the sensor's output impedance. If the sensor is capacitive, its output capacitance will be very quickly charged by the leakage current. This will not only cause an error, but may even lead to the sensor's destruction, especially if the sensor uses some chemical compound (e.g., a resistive moisture sensor).

There are several techniques known to minimize the board leakage current effect. One is a careful board layout to keep higher voltage conductors away from the high-impedance components. A leakage through the board thickness in multilayer boards should not be overlooked. Another method is electrical guarding, which is an old trick. The so-called driven shield is also highly effective. Here, the input circuit is surrounded by a conductive trace that is connected to a low-impedance point at the same potential as the input. The guard absorbs the leakage from other points on the board, drastically reducing currents that may reach the input terminal. To be completely effective, there should be guard rings on both sides of the printed circuit board. As an example, an amplifier is shown with a guard ring, driven by a relatively low impedance of the amplifier's inverting input.

It is highly advisable to locate the high-impedance interface circuits as close as possible to the sensors. However, sometimes connecting lines can not be avoided. Coaxial shielded cables with good isolation are recommended [1]. Polyethylene or virgin (not reconstructed) Teflon is best for the critical applications. In addition to potential insulation problems, even short cable runs can reduce bandwidth unacceptably with high source resistances. These problems can be largely avoided by bootstrapping the cable's shield. Figure 5.4b shows a voltage follower connected to the inverting input of an amplifier. The follower drives the shield of the cable, thus reducing the cable capacitance, the leakage and spurious voltages resulting from cable flexing. A small capacitance at the follower's noninverting input improves its stability.

Another problem that must be avoided is connecting to the input of an amplifier any components, besides a sensor, that potentially may cause problems. An example

of such a "troublemaker" is a ceramic capacitor. In a hope to filter out high-frequency transmitted noise at the input, a designer quite frequently uses filter capacitors either at the input, or in the feedback circuit, of an input stage. If for a cost-saving or space saving reason she selects a ceramic capacitor, she may get what is not expecting. Many capacitors possess the so-called dielectric absorption properties, which are manifested as a memory effect. If such a capacitor is subjected to a charge spike either from a sensor, or from a power supply, or just from any external noise source, the charge will alter the capacitor's dielectric properties in such a way as a capacitor now behaves like a small battery. That "battery" may take a long time to lose its charge: from few seconds to many hours. The voltage generated by that "battery" is added to the sensor's signal and may cause significant errors. If a capacitor must be employed at the input stage, a film capacitor should be used instead of ceramic.

5.2 Amplifiers

Most passive sensors produce weak output signals. The magnitudes of these signals may be on the order of microvolts (μV) or picoamperes (pA). On the other hand, standard electronic data processors, such as A/D converters, frequency modulators, data recorders, etc. require input signals of sizable magnitudes on the order of volts (V) and milliamperes (mA). Therefore, an amplification of the sensor output signals has to be made with a voltage gain up to 10,000 and a current gain up to 1 million. Amplification is part of a signal conditioning. There are several standard configurations of amplifiers that might be useful for the amplifying signals from various sensors. These amplifiers may be built of discrete components, such as semiconductors, resistors, capacitors, and inductors. Alternatively, the amplifiers are frequently composed of standard building blocks, such as operational amplifiers and various discrete components.

It should be clearly understood that a purpose of an amplifier is much broader than just increasing the signal magnitude. An amplifier may be also an impedance-matching device, an enhancer of a signal-to-noise ratio, a filter, and an isolator between input and output.

5.2.1 Operational Amplifiers

One of the principle building blocks for the amplifiers is the so-called *operational amplifier* (OPAM), which is either an integrated (monolithic) or hybrid (a combination of monolithic and discrete parts) circuit. An integrated OPAM may contain hundreds of transistors, as well as resistors and capacitors. An analog circuit designer, by arranging around the OPAM discrete components (resistors, capacitors, inductors, etc.), may create an infinite number of useful circuits, not only the

amplifiers, but many others circuits as well. Operational amplifiers are also used as cells in custom-made integrated circuits of the analog or mixed technology types. These circuits are called *application-specific integrated circuits* or ASICs for short. Below, we will describe some typical circuits with OPAM, which are often used in conjunction with various sensors.

As a building block, a good operational amplifier has the following properties (a schematic representation of OPAM is shown in Fig. 5.5):

- Two inputs: one is inverting (−) and the other is noninverting (+);
- A high input resistance (on the order of hundreds of MΩ or even GΩ);
- A low output resistance (a fraction of Ω);
- An ability to drive capacitive loads;
- A low input offset voltage e_o (few mV or even μV);
- A low input bias current i_o (few pA or even less);
- a very high *open loop gain* A_{OL} (at least 10^4 and preferably over 10^6). That is, the OPAM must be able to magnify (amplify) a voltage difference V_{in}, between its two inputs by a factor of A_{OL};
- a high common mode rejection ratio (CMRR). That is, the amplifier suppresses the in-phase equal magnitude input signals (common-mode signals) V_{CM} applied to its both inputs;
- low intrinsic noise;
- a broad operating frequency range;
- a low sensitivity to variations in the power supply voltage;
- a high environmental stability of its own characteristics.

For the detailed information and the application guidance the user should refer to data sheets and catalogues published by the respective manufacturers. Such catalogues usually contain the selection guides for every important feature of an OPAM. For instance, OPAMs are grouped by such criteria as low offset voltages, low bias currents, low noise, bandwidth, etc.

Figure 5.5a depicts an operational amplifier without any feedback components. Therefore, it operates under the so-called open-loop conditions. An open loop gain,

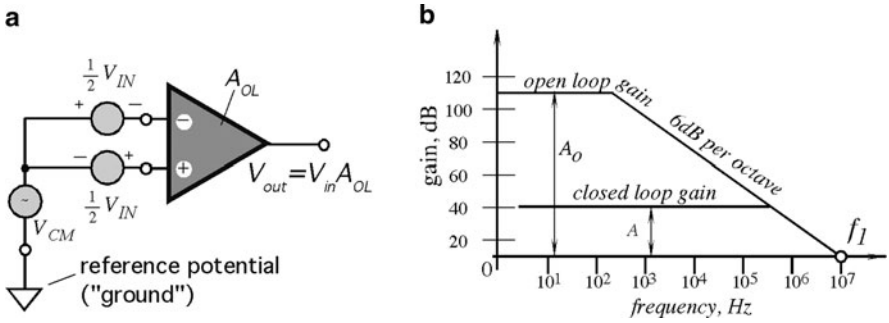

Fig. 5.5 General symbol of an operational amplifier (**a**), and gain/frequency characteristic of an OPAM (**b**)

A_{OL}, of an OPAM is always specified but is not a very stable parameter. Its frequency dependence may be approximated by a graph of Fig. 5.5b. The A_{OL} changes with the load resistance, temperature, and the power supply fluctuations. Many amplifiers have an open loop gain temperature coefficient on the order of 0.2 to 1%/°C and the power supply gain sensitivity on the order of 1%/%. An OPAM is very rarely used with an open loop (without the feedback components) because the high open-loop gain may result in circuit instability, a strong temperature drift, noise, etc. For instance, if the open-loop gain is 10^5, the input voltage drift of 10 μV would cause the output drifts by about 1 V.

The ability of an OPAM to amplify small-magnitude high-frequency signals is specified by the gain-bandwidth product (GBW), which is equal to the frequency f_1 where the amplifier gain becomes equal to unity. In other words, above the f_1 frequency, the amplifier cannot amplify. Figure 5.6a depicts a noninverting amplifier where resistors R_1 and R_2 define the feedback loop. The resulting gain $A = 1 + R_2/R_1$ is a closed-loop gain. It may be considered constant over a much broader frequency range [see Fig. 5.5b], however, f_1 is the frequency limiting factor regardless of the feedback. A linearity, gain stability, the output impedance, and gain accuracy are all improved by the amount of feedback and now depend mainly on characteristics of the feedback components. As a general rule for moderate accuracy, the open loop gain of an OPAM should be at least 100 times greater than the closed loop gain at the highest frequency of interest. For even higher accuracy, the ratio of the open and closed loop gains should be 1,000 or more.

A typical data sheet for an OPAM specifies the bias and offset voltages. Due to limitations in manufacturing technologies, any OPAM acts not only as a pure amplifier, but as a generator of voltages and currents, which may be related to its input [Fig. 5.6b]. Since these spurious signals are virtually applied to the input terminals, they are amplified along with the useful signals.

Because of offset voltages and bias currents, an interface circuit does not produce zero output when zero input signal is applied. In dc-coupled circuits, these undesirable input signals may be indistinguishable from the useful signal. If the input offset voltage is still too large for the desired accuracy, it can be trimmed

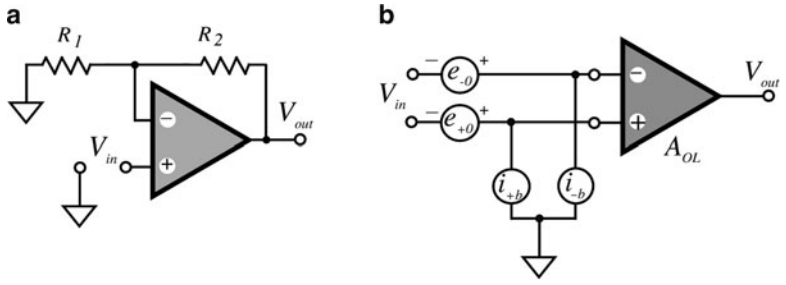

Fig. 5.6 Noninverting amplifier (**a**); offset voltages and bias currents in an operational amplifier are represented by generators connected to its inputs (**b**)

out either directly at the amplifier (if the amplifier has dedicated trimming terminals) or in the independent offset compensation circuit.

An application engineer should be concerned with, the output offset voltage, which can be derived from formula:

$$V_o = A(e_o + i_o R_{eqv}) \tag{5.7}$$

where R_{eqv} is the equivalent resistance at the input (a combination of the sensor's output resistance and the input resistance of the amplifier), e_o is the input offset voltage, and i_o is the input bias current. The offset is temperature-dependent. In circuits where the amplifier has high gain, the output voltage offset may be a source of substantial error. There are several ways to handle this difficulty. Among them is selecting an amplifier with low bias current, high input resistance, and low offset voltage. Chopper-stabilized amplifiers are especially efficient for reduction of offset voltages.

5.2.2 Voltage Follower

A voltage follower (Fig. 5.7) is a an electronic circuit that provides impedance conversion from a high to low level. A typical follower has high input impedance (the high input resistance and the low input capacitance) and low output resistance (the output capacitance makes no difference). A good follower has a voltage gain very close to unity (typically, 0.999 at lower frequencies) and a high current gain. In essence, it is a current amplifier and impedance converter. Its high input and low output impedances make it indispensable for interfacing between many sensors and signal processing devices.

A follower, when connected to a sensor, makes very little effect on the latter's performance, thus providing a buffering function between the sensor and the load. When designing a follower, these tips might be useful:

- For the current-generating sensors, the input bias current of the follower must be at least 100 times smaller than the sensor's current.
- The input offset voltage must be either trimmable or smaller than the required LSB.
- The temperature coefficient of the bias current and the offset voltage should not result in errors of more than 1 LSB over an entire temperature range.

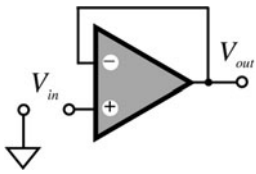

Fig. 5.7 Voltage follower with an operational amplifier

5.2.3 *Instrumentation Amplifier*

An instrumentation amplifier (IA) has two inputs and one output (Fig. 5.8). It is distinguished from an operational amplifier by its finite gain (which is usually no more than 100) and the availability of both inputs for connecting to the signal sources. The latter feature means that all necessary feedback components are connected to other parts of the instrumentation amplifier, rather than to its non-inverting and inverting inputs. The main function of the IA is to produce an output signal which is proportional to the difference in voltages between its two inputs:

$$V_{out} = a(V_+ - V_-) = a\Delta V, \tag{5.8}$$

where V_+ and V_- are the input voltages at noninverting and inverting inputs respectively, and a is the gain. It is important to assure high input resistances for both inputs, so that the amplifier can be used in a true differential form. A differential input of the amplifier is very important for rejection of common mode interferences having an additive nature (see Sect. 5.10). Thus, the IA should have a high common-mode rejection ratio (CMRR), that is, its output signal should be insensitive to the value of V_+ or V_- but responsive only to their difference.

An instrumentation amplifier can be either built from several discrete OPAMs, in a monolithic or hybrid forms. Presently, instrumentation amplifiers are readily available from many manufacturers in form of monolithic chips. An example of a high-quality monolithic instrumentation amplifier is INA118 from Burr-Brown/Texas Instruments (www.ti.com). It offers low offset voltage of 50 μV (the input signal produced by the IA when both inputs are connected together) and high CMRR (110 dB). The gain is programmed by a single resistor. Also, many integrated circuits, such as microcontrollers or DSP (digital signal processors) have built-in input instrumentation amplifiers for direct interface between the input sensors and the internal A/D (analog-to-digital) converters.

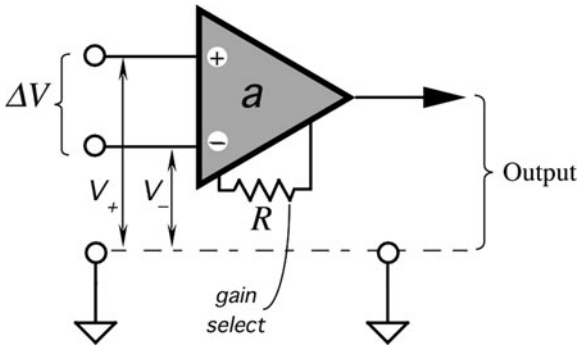

Fig. 5.8 Instrumentation amplifier

5.2.4 Charge Amplifiers

Charge amplifiers (CA) is a very special class of circuits, which must have extremely low bias currents. These amplifiers are employed to convert to voltage signals from the capacitive sensors, quantum detectors, pyroelectric sensors, and other devices, which generate very small charges (on the order of pico-coulombs, pC) or currents (on the order of pico-amperes). A basic circuit of a charge-to-voltage converter is shown in Fig. 5.9a. A capacitor, C, is connected into a feedback network of an OPAM. Its leakage resistance r must be substantially larger than the impedance of the capacitor at the lowest operating frequency. A good film capacitor is usually recommended along with a good quality printed circuit board where the components are coated with conformal coating.

A transfer function of the converter is

$$V_{out} = -\frac{\Delta Q}{C}. \tag{5.9}$$

Special integrated charge sensitive preamplifiers are commercially available for precision applications.

Many sensors can be modeled as capacitors. Some capacitive sensors are active, that is, they require an excitation signal. Examples are the microphones, capacitive force, and pressure transducers and humidity detectors. Other capacitive sensors are passive, that is they directly convert a stimulus into an electric charge or current. Examples are the piezoelectric and pyroelectric detectors. There are also noncapacitive sensors that still can be considered as current generators. An example is a photodiode.

A current generating sensor is modeled by a leakage resistance, r, connected in parallel with a current generator that has an infinitely high internal resistance (Fig. 5.10). The sensor generates current, i, which has two ways to outflow: to the sensors leakage resistance, r, as current, i_o, and the other, i_{out}, toward the interface circuit input impedance, Z_L. Naturally, current i_o is useless and to minimize the

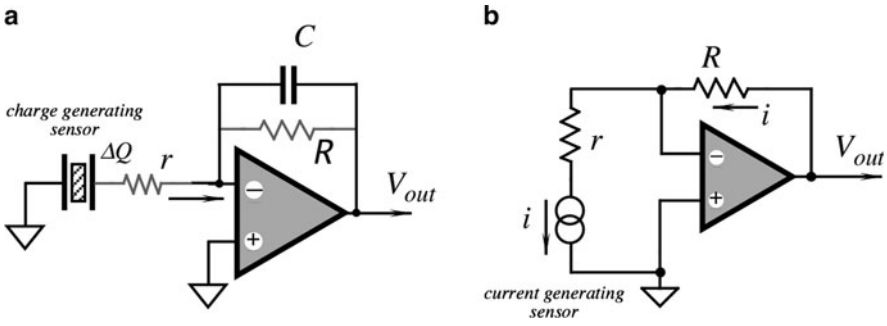

Fig. 5.9 Charge-to-voltage (**a**) and current-to-voltage (**b**) converters

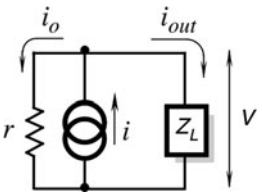

Fig. 5.10 An equivalent circuit of a current-generating sensor

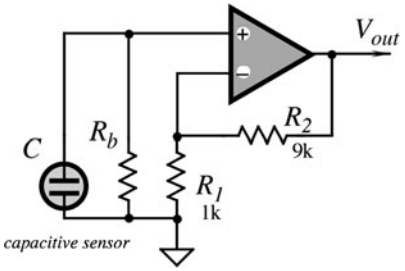

Fig. 5.11 Noninverting current-to-voltage converter

error of the current-to-voltage conversion, the leakage resistance of the sensor must be much larger than the impedance of the interface circuit (Fig. 5.11).

Ohm's law suggests that to convert electric current i_{out} into voltage, current should pass through an appropriate impedance and the voltage drop across that impedance is proportional to the magnitude of the current. Figure 5.9b shows a basic current-to-voltage converter where the current-generating sensor is connected to the inverting input of an OPAM, which serves as a virtual ground. In other words, voltage at the inverting input is almost equal to that at the noninverting input, which is grounded. The sensor operates at nearly zero voltage across its terminals and its current defines the output voltage of the OPAM:

$$V_{out} = -iR \tag{5.10}$$

Resistor, $r << R$ is often required for the circuit stability because at high frequencies, without that resistor the OPAM would operate near the open loop gain, which may result in oscillations. This is especially true than the sensor has small internal resistance. The advantage of the virtual ground is that the output signal does not depend on the sensor's capacitance. The circuit produces voltage whose phase is shifted by 180° with respect to the current. A noninverting circuit shown in Fig. 5.12a can convert and amplify the signal, however, its speed response depends on both the sensor's capacitance and the converting resistor R_1. Thus, the response to a step function in a time domain can be described by

$$V_{out} = iR_b\left(1 + \frac{R_2}{R_1}\right)(1 - \mathrm{e}^{-\frac{t}{rC}}). \tag{5.11}$$

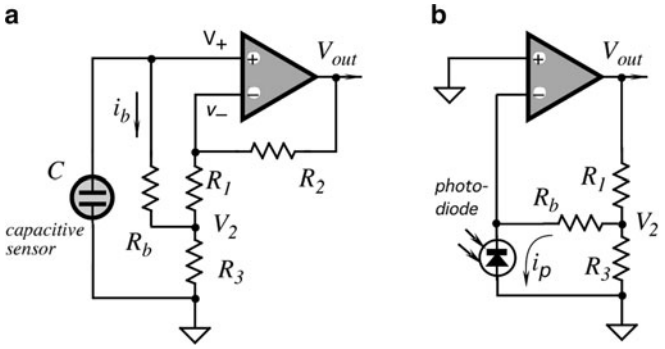

Fig. 5.12 Resistance multipliers (**a**) and (**b**)

When converting currents from such sensors as piezo and pyroelectrics, the resistor R_b [R in circuit 5.9(B)] may be required on the order of tens or even hundreds of gigohms. In many cases, resistors of such high values may be not available or impractical to use due to poor environmental stability. A high ohmic resistor can be simulated by a circuit, which is known as a resistance multiplier. For a positive input of an amplifier is implemented by adding a positive feedback around the amplifier. Figure 5.12a shows that R_1 and R_3 form a resistive divider. Due to a high open loop gain of the OPAM, voltages at noninverting and inverting inputs are almost equal to one another: $V_+ \approx V_-$. As a result, voltage, V_2, at the divider is

$$V_2 = V_- \frac{R_3}{R_1 + R_3} \approx V_+ \frac{R_3}{R_1 + R_3}, \tag{5.12}$$

and current through the resistor is defined trough the voltage drop:

$$i_b = \frac{\Delta V}{R_b} = \frac{V_+ - V_2}{R_b} = \frac{V_+}{R_b} \frac{R_1}{R_1 + R_3}. \tag{5.13}$$

From this equation, the input voltage can be found as a function of the input current and the resistive network:

$$V_+ = i_b R_b \left(1 + \frac{R_3}{R_1}\right). \tag{5.14}$$

It's seen that the resistor R_b is multiplied by a factor of $(1 + R_3/R_1)$. For example, if the highest resistor you may consider is 10 MΩ, by selecting the multiplication factor in parenthesis of say 5, you'll get a virtual resistance of 50 MΩ. Resistance multiplication, while being a powerful trick, should be used with caution. Specifically, noise, bias current and offset voltage, all of them are also multiplied by the

same factor $(1 + R_3/R_1)$, which may be undesirable in some applications. Further, since the network forms a positive feedback, it may cause circuit instability. Therefore, in practical circuits, a resistance multiplication should be limited to a factor of 10. If a resistance multiplication is required for a negative input of an amplifier, the circuit of Fig. 5.12b comes in handy. the resistance multiplication is governed by the same (5.14).

5.3 Light-to-Voltage Converters

Light-to-voltage converters are based on combination of photosensors and current-to-voltage converter circuits. For detecting extremely low-intensity light, typically one or several photons, the photomultipliers are generally employed (Chap. 15), however, for less demanding applications three types of a photosensor are available: a photodiode, phototransistor, and photoresistor (Chap. 14). They employ a photoeffect that was discovered by Albert Einstein and won him the Nobel Prize. The difference between a photodiode and a phototransistor is in construction of the semiconductor chip. A photodiode has one p-n junction, while a phototransistor has two junctions where the base of the transistor may be floating or may have a separate terminal. The base current is a photo-induced current that is multiplied by the transistor's β to produce the collector current. Thus, a phototransistor is equivalent to a photodiode with a built-in current amplifier.

From the electrical point of view, a photodiode can be represented by an equivalent circuit shown in Fig. 5.13a. It consists of a current generator (internal input impedance is infinitely large), a parallel regular diode (like a rectifier diode), resistance of the junction R_j, capacitance of the junction C_j, and a serial resistance R_s. The current generator generates a photocurrent proportional to the photon flux. This current flows in the direction from the cathode (−) to the anode (+) of the photodiode. Note that for very strong illuminations, the photocurrent will start flowing through a nonlinear rectifier diode, which will degrade a linearity.

A photodiode can be used in voltaic or current modes. In the voltaic mode, a photodiode is connected to a very high resistor (10^7–10^9 Ω) and a good voltage

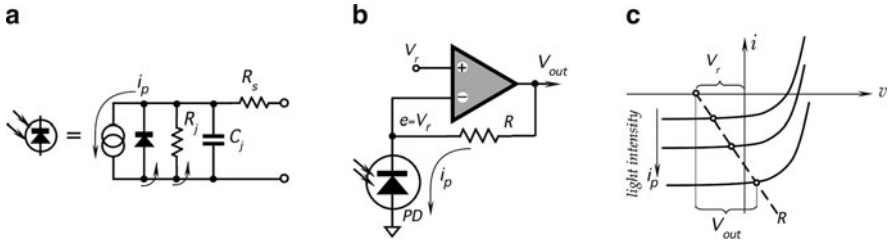

Fig. 5.13 Equivalent circuit of a photo-diode (**a**) A reverse-biased photo-diode with a current-to voltage converter; (**b**) load diagram of the circuit (**c**)

amplifier. The diode will work like a battery with voltage proportional to the light intensity. This voltage is the result of a photocurrent i_p passing through the internal junction resistance R_j. In a current mode, the photodiode is virtually shorted (a voltage across the diode is zero) and current i_p is drawn to the current-to-voltage converter as described below. This node is more popular, especially for applications where a high-speed response is required.

A circuit with an operational amplifier is shown in Fig. 5.13b. Note that the reference voltage V_r creates a constant reverse bias across the photodiode. Figure 5.13c shows the operating points for a load resistor R. Features of the circuits used with a reverse-biased photodiode are high-speed response and wide-proportional-range of output. Therefore, this circuit is generally used. Another circuit with an operational amplifier has a zero-bias across the photodiode as shown in Fig. 5.14a. This arrangement provides a near-ideal short-circuit current in a wide operating range. The output voltage (V_{OUT}) is given as $V_{OUT} = i_p R$. Figure 5.14b shows the output voltage vs. radiant intensity (a transfer function). An arrangement with no bias and high-impedance loading to the photodiode provides less influence by dark current and a wide linear range of the photocurrent relative to the radiant intensity. It should be noted that for rather small illuminations to get a meaningful output (few hundred millivolts), a value of resistor R should be quite large on the order of 100 MΩ or even several GΩ. If such resistors are not available, a resistance multiplication circuit as shown in Fig. 5.12c may be cautiously employed. The multiplication factor should not be greater than 10 as all the bad things in the circuit are also multiplied: the offset voltage, bias current, and noise. When using the high-Ohmic resistors, the photosensor and interface circuit should be electrically shielded. Even a minute capacitive coupling of the environment to such resistors brings in a lot of interferences, especially from power lines (60 or 50 Hz).

The interface circuits for a phototransistor are similar, except that they have to provide a voltage across the collector-emitter terminals as shown in Fig. 5.15a. The transfer function of this circuit is shown in Fig. 5.15b. A phototransistor circuit is more sensitive to light but for the price of higher nonlinearity at stronger irradiances.

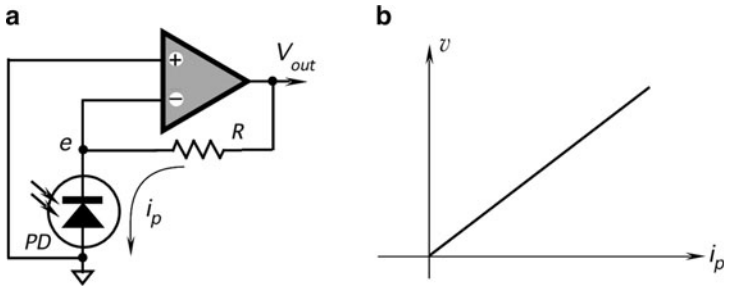

Fig. 5.14 A zero-biased photodiode with a current-to-voltage converter (**a**) and a transfer function (**b**)

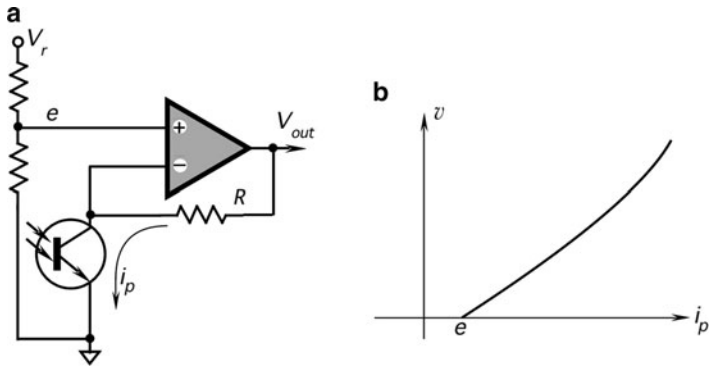

Fig. 5.15 Light-to-voltage converted with a photo-transistor (**a**); transfer function (**b**)

We are not describing here interface circuits for photoresistors because any suitable resistance measurement circuits can be used for the purpose. An example is the Wheatstone bridge circuits which we will discuss below.

5.4 Excitation Circuits

External power is required for operation of the *active* sensors. Examples are: temperature sensors (thermistors and RTDs), pressure sensors (piezoresistive and capacitive), and displacement (electromagnetic and resistive). The power may be delivered to a sensor in different forms. It can be a constant voltage, constant current, and sinusoidal or pulsing currents. It may even be delivered in the form of light or ionizing radiation. The name for that external power is an excitation signal. In many cases, stability and precision of the excitation signal directly relates to the sensor's accuracy and stability. Hence, it is imperative to generate the signal with such accuracy that the overall performance of the sensing system is not degraded. Below, we review several electronic circuits that feed sensors with appropriate excitation signals.

5.4.1 Current Generators

Current generators are often used as excitation circuits to feed sensors with predetermined currents that, within limits, are independent of the sensor properties, stimulus value, or environmental factors. In general terms, a current generator (current pump) is a device which produces electric current independent of the load impedance. That is, within the capabilities of the generator, amplitude of its output current must remain substantially independent of any changes in the

impedance of the load. It is said that an ideal current source (generator) has infinitely high output resistance, so any series load value will not change anything. When generating current for a variable load, according to Ohm's law, the corresponding voltage must change in synch.

Usefulness of the current generators for the sensor interfaces is in their ability to produce excitation currents of precisely controlled magnitude and shape. Hence, a current generator should not only produce current which is load independent, but it also must be controllable from an external signal source (a wave-form generator), which in most cases has a voltage output. A good current generator must produce current that follows the control signal with high fidelity and is independent of the load over a broad range of impedances.

There are two main characteristics of a current generator: the output resistance and the voltage compliance. The output resistance should be as high as practical. A voltage compliance is the highest voltage that can be developed across the load without affecting the output current. For a high resistive load, according to Ohm's law, a higher voltage is required for a given current. For instance, if the required excitation current is $i = 10$ mA and the highest load impedance at any given frequency is Z_L=10 kΩ, a voltage compliance of at least $iZ_L = 100$ V would be needed. Below, we cover some useful circuits with increased voltage compliance where the output currents can be controlled by external signals.

A unipolar current generator is called either a current source (generates the outflowing current), or a current sink (generates the in-flowing currents). Here, unipolar means that it can produce currents flowing in one direction only, usually toward the ground. Many of such generators utilize current-to-voltage characteristics of transistors. A voltage-controlled current source or sink may include an operational amplifier (Fig. 5.16). In such a circuit, a precision and stable resistor R_1 defines the output current, i_{out}. The circuit contains a feedback loop through the OPAM that keeps voltage across resistor R_1 constant and thus the constant current. To deliver a higher current at a maximum voltage compliance, a voltage drop across the sensing resistor R_1 should be as little as possible. In effect, that current is equal to V_1/R_1. For better performance, the current through the base of the output transistor should be minimized, hence, a field-effect rather than bipolar transistor is often used as an output current delivering device.

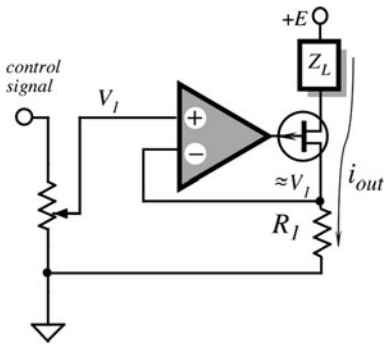

Fig. 5.16 Current sources with OPAM

It is well known hat the transistor's collector current is very little dependent on collector voltages. This feature was employed by the so-called current mirrors. A current mirror has one current input and at least one (may be several) current output. Therefore, the output current is controlled by the input current. The input current is supplied from an external source (like a voltage source plus a resistor) and should be of a known value. The so-called Wilson current mirrors have the control currents of about the same magnitude as the output currents, that is, they have an input-output ratio 1:1. Mirrors with other ratios also were designed, such as 1:2 and 1:4. Commercially available current mirrors may have a very broad current range, such as from 3 nA to 3 mA as in the integrated current mirror ADL5315 from Analog Devices.

For many sensors, bipolar current generators may be required. Such a generator provides a sensor with the excitation current, which may flow in both directions (in- and out-flowing). Figure 5.17 shows a noninverting (B) and inverting (A) circuits with an operational amplifier where the load is connected as a feedback. Current though the load Z_L, is equal to V_1/R_1 and is load-independent. The load current follows V_1 within the operating limits of the amplifier. An obvious limitation of the circuit is that the load is "floating," i.e., it is not connected to a ground bus or any other reference potential. For some applications, this is quite all right, however, many sensors need to be grounded or otherwise referenced. A circuit shown in Fig. 5.17b keeps one side of the load impedance near the ground potential, because a noninverting input of the OPAM is a virtual ground. Nevertheless, even in this circuit, the load is still fully isolated from the ground. One negative implication of this isolation may be an increased pick up of various kinds of transmitted noise.

In cases where the sensor must be grounded, a current pump invented by Brad Howland at MIT may be used [Fig. 5.18a]. The pump operation is based on utilizing both negative and positive feedbacks around the operational amplifier. The load is connected to the positive loop [2]. Current through the load is defined by

$$i_{out} = \frac{R_2}{R_1}\frac{(V_1 - V_2)}{R_5} \tag{5.15}$$

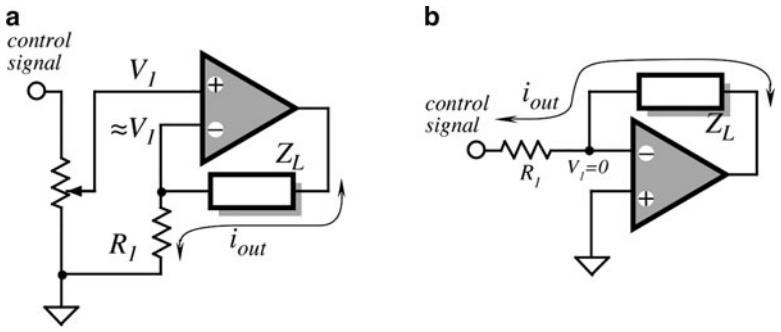

Fig. 5.17 Bipolar current generators with floating loads noninverting circuit (**a**); circuit with a virtual ground (**b**)

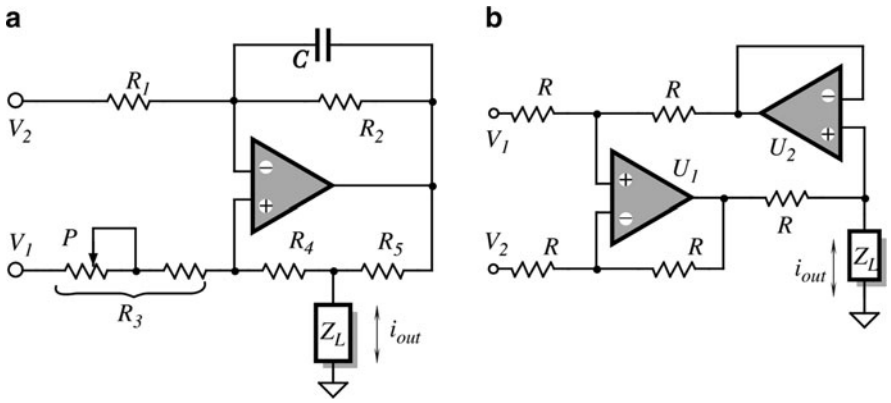

Fig. 5.18 Current generators with ground referenced loads Howland current pump (**a**); current pump with two OPAMs (**b**)

A trimming resistor, P, must be adjusted to assure that

$$R_3 = R_1 \frac{R_4 + R_5}{R_2} \tag{5.16}$$

In that circuit, each resistor may have a relatively high value (100 kΩ or higher), but the value for R_5 should be relatively small. This condition improves efficiency of the Howland current pump, as smaller voltage is wasted across R_5 and smaller current is wasted through R_4 and R_3. The circuit is stable for most of the resistive loads, however, to insure stability, a few picofarad capacitor C may be added in a negative feedback or/and from the positive input of an operational amplifier to ground. When the load is inductive, an infinitely large compliance voltage would be required to deliver the set current when a fast transient control signal is applied. Therefore, the current pump will produce a limited rising slope of the output current. The flowing current will generate an inductive spike across the output terminal, which may be fatal to the operational amplifier. It is advisable, for the large inductive load to clamp the load with diodes to the power supply buses.

An efficient current pump with four matched resistors and two operational amplifiers is shown in Fig. 5.18b. Its output current is defined by the equation

$$i_{out} = \frac{(V_1 - V_2)}{R_s}. \tag{5.17}$$

The advantage of this circuit is that resistors R may be selected with a relatively high value and housed in the same thermally homogeneous packaging for better thermal tracking.

5.4.2 Voltage References

A voltage reference is an electronic device that generates constant voltage that is affected little by variations in power supply, temperature, load, aging, and other factors. There are several techniques known for generation of such voltages. Many voltage references are available in monolithic form for a large variety of voltages. Most of them operate with the so-called band gap references.

5.4.3 Oscillators

Oscillators are generators of variable electrical signals. In many applications that employ microprocessors or microcontrollers, square-wave pulses may be generated at one of the I/O ports. When no such port is available, free-standing oscillators may be developed. Any oscillator is essentially comprised of a circuit with a gain stage and some nonlinearity, and a certain amount of positive feedback. By definition, an oscillator is an unstable circuit (as opposed to an amplifier which better be stable!) whose timing characteristics should be either steady or changeable according to a predetermined functional dependence. The latter is called a *modulation*. Generally, there are three types of electronic oscillators classified according to the time-keeping components: the *RC*, *LC*, and crystal oscillators. In the *RC* oscillators, the operating frequency is defined by capacitors (*C*) and resistors (*R*), in the *LC*-oscillators by the capacitive (*C*) and inductive (*L*) components. In the crystal oscillators, operating frequency is defined by a mechanical resonant in specific cuts of piezoelectric crystals, usually quartz and ceramics. There is a great variety of oscillation circuits, coverage of which is beyond the scope of this book. Below, we briefly describe some practical circuits.

Many various multivibrators can be built with logic circuits, for instance with NOR, NAND gates, or binary inverters. Also, many multivibrators can be designed with comparators or operational amplifiers having a high open loop gain. In all these oscillators, a combination of a capacitor and a resistor is time-keeping combination. These circuits are called relaxation oscillators. A voltage across a charged capacitor is compared with another voltage, which is either constant or changing with a different rate. The moment when both voltages are equal is detected by a comparator. The indication of a comparison is fed back to the *RC*-network to alter the capacitor charging in the opposite direction, which is discharging. Recharging in a new direction goes on until the next moment of comparison. This basic principle essentially requires the following minimum components: a capacitor, a charging circuit, and a threshold device (a comparator). Several monolithic relaxation oscillators are available from many manufacturers, for instance a very popular timer, type 555, which can operate in either monostable or actable modes. For the illustration, below we describe just two discrete-component square-wave oscillators, however, there is a great variety of such circuits, which the reader can find in many books on operational amplifiers and digital systems, for instance [3].

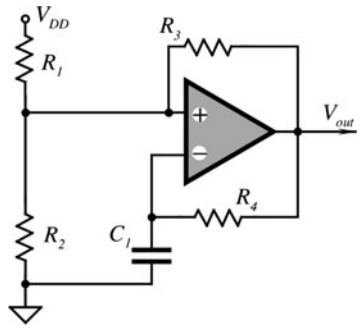

Fig. 5.19 Square-wave oscillator with OPAM

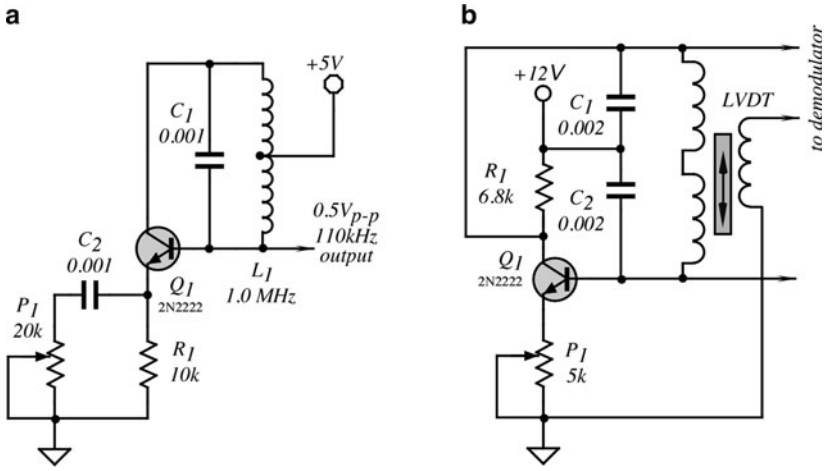

Fig. 5.20 *LC* sine-wave oscillators

A very popular square-wave oscillator (Fig. 5.19) can be built with one OPAM or a voltage comparator.[1] The amplifier is surrounded by two feedback loops: one is negative (to an inverting input) and the other is positive (to a noninverting input). The positive feedback (via R_3) controls the threshold level, while the negative loop charges and discharges timing capacitor C_1, through the resistor R_4. The frequency of this oscillator can be determined from

$$f = \frac{1}{R_4 C_1}\left[\ln\left(1 + \frac{R_1||R_2}{R_3}\right)\right]^{-1}, \tag{5.18}$$

where $R_1||R_2$ is an equivalent resistance of parallel-connected R_1 and R_2.

[1]A voltage comparator differs from an operational amplifier by its faster speed response and the output circuit which is easier interfaceable with digital circuits.

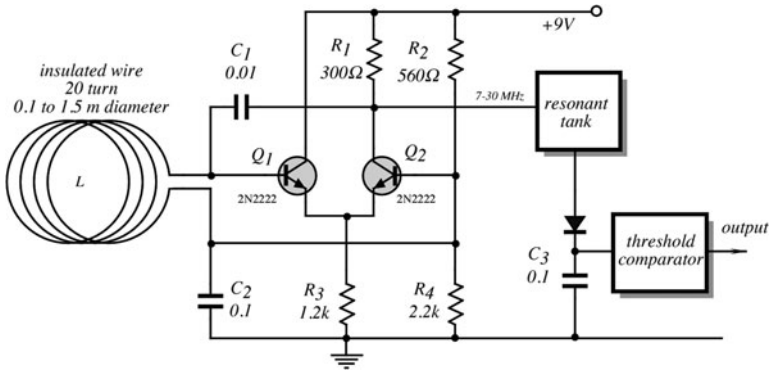

Fig. 5.21 *LC* radio-frequency oscillator as a capacitive occupancy detector

Two circuits shown in Fig. 5.20 can generate sine wave signals. They use the *npn*-transistors as amplifiers and the *LC*-networks to set the oscillating frequency. The (b) circuit is especially useful for driving the LVDT position sensors, as the sensor's transformer becomes a part of the oscillating circuit.

A radiofrequency oscillator can be used as part of a capacitive occupancy detector to detect presence of people in the vicinity of its antenna (Fig. 5.21).[2] The antenna is a coil, which together with the C_2 capacitor, determines the oscillating frequency. The antenna is coupled to the environment through its distributed capacitance, which somewhat reduces the frequency of the oscillator. When a people move into vicinity of the antenna, they bring in an additional capacitance that lowers the oscillator frequency even further. The output of the oscillator is coupled to a resonant tank (typically, an *LC*-network) which is tuned to a baseline frequency (near 30 MHz).

5.4.4 Drivers

As opposed to current generators, voltage drivers must produce output voltages, which over broad ranges of the loads and operating frequencies are independent of the output currents. Sometimes, the voltage drivers are called *hard voltage sources*. Usually, when the sensor, which has to be driven is purely resistive, a driver can be a simple output stage, which can deliver sufficient current. However, when the load contains capacitances or inductances, that it, the load is reactive, the output stage becomes a more complex device.

In many instances, when the load is purely resistive, there still can be some capacitance associated with it. This may happen when the load is connected though lengthy wires or coaxial cables. A coaxial cable behaves as a capacitor connected from its central conductor to its shield if the length of the cable is less than 1/4 of the

[2]See Sect. 6.3.

wavelength in the cable at the frequency of interest f. For a coaxial cable, this maximum length is given by

$$L \leq 0.0165\frac{c}{f}, \tag{5.19}$$

where c is the velocity of light in a coaxial cable dielectric.

For instance, if $f = 100$ kHz, $L \leq 0.0165\frac{3\times10^8}{10^5} = 49.5$, that is, a cable less than 49.5 m (162.4 ft) long will behave as a capacitor connected in parallel with the load [Fig. 5.22a]. For an R6-58A/U cable, the capacitance is 95 pF/m. This capacitance must be considered for two reasons: for the speed and stability of the circuits. The instability results from the phase shift produced by the output resistance of the driver R_o and the loading capacitance C_L:

$$\varphi = \text{arctang}(2\pi f R_o C_L). \tag{5.20}$$

For example, if $R_o = 100\ \Omega$ and $C_L = 1{,}000$ pF, at $f = 1$ MHz, the phase shift $\varphi \approx 32^\circ$. This shift significantly reduces the phase margin in a feedback network, which may cause a substantial degradation of the response and a reduced ability to drive capacitive loads. The instability may be either overall, when an entire system oscillates, or localized when the driver alone becomes unstable. The local instabilities often can be cured by large bypass capacitors (on the order of 10 μF) across the power supply or the so-called Q-spoilers consisting of a serial connection of 3–10 Ω resistor and a disk ceramic capacitor connected from the power supply pins of the driver chip to ground.

To make a driver stage more tolerant to capacitive loads, it can be isolated by a small serial resistor as it is shown in Fig. 5.22b. A small capacitive feedback (C_f) to the inverting input of the amplifier, and a 10 Ω resistor may allow to drive loads as large as 0.5 μF. However, in any particular case, it is recommended to find the best values for the resistor and the capacitor experimentally.

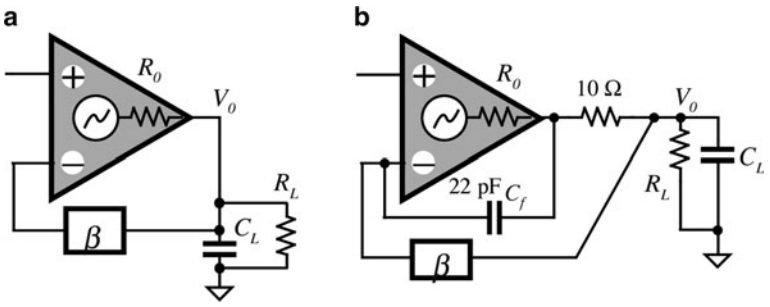

Fig. 5.22 Driving a capacitive load load capacitor is coupled to the driver's input through a feedback (**a**); decoupling of a capacitive load (**b**)

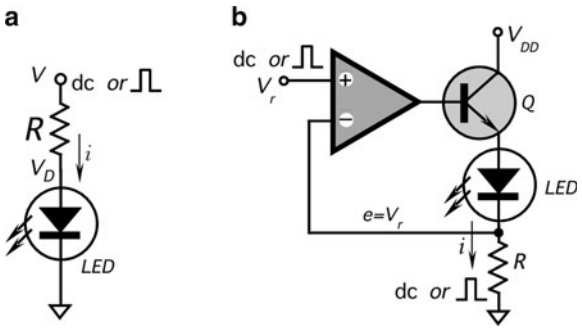

Fig. 5.23 Resistive (**a**) and currents source (**b**) drivers for LED

5.4.5 Optical Drivers

In same applications, optical sensors receive light from natural sources, like the celestial objects or fire, or from human-made sources like scintillating materials, incandescent or fluorescent light. In many other cases, a special light source must be provided. An example is a light detector in a TV set remote control receiver. This sensor produces no output till it receives a train of light (near infrared) pulses from the remote control transmitter. The transmitter must contain a light emitter with the matching spectral characteristics. The most popular light emitters are light emitting diodes (LED) that operate from the UV to near infrared spectral range. The LED produces light whose light intensity is nearly proportional to the current passing through the diode. The simplest drive circuit for LED is shown in Fig. 5.23a. It contains either a dc voltage source or a pulsing voltage source and a current limiting resistor R. Current i is defined by formula

$$i = \frac{V - V_D}{R}, \tag{5.21}$$

where V_D is voltage across the LED (typically from 1.5 to 2.0 V). This voltage depends on the current and temperature and, subsequently the intensity of light generated by the LED also will be the current- and temperature-dependent. For precision applications, the current through LED should be maintained constant. Thus, instead of a current limiting resistor, an electronic current source should be employed as shown in Fig. 5.23b. In this driver, the current is solely defined by the drive control voltage and sensing resistor R and thus is temperature-independent.

5.5 Analog-to-Digital Converters

5.5.1 Basic Concepts

The analog-to-digital converters (abbreviated as A/D, or ADC, or A2D, or A-to-D) range from discrete circuits, to monolithic ICs (integrated circuits), to high-performance hybrid circuits, modules, and even boxes. Also, the converters are available as

standard cells for custom and semicustom application-specific integrated circuits (ASIC). The A/D converters transform analog data, usually voltage, into an equivalent digital format, compatible with digital data processing devices. Key characteristics of the A/D converters include absolute and relative accuracy, linearity, no-missing codes, resolution, conversion speed, stability, and price. Quite often, when price is of a major concern, the monolithic IC versions (integrated circuits) are the most efficient. The most popular A/D converters are based on a successive approximation technique because of an inherently good compromise between speed and accuracy. However, other popular techniques are used in a large variety of applications, especially when no high conversion speed is required. These include dual-ramp, quad-slope, pulse-width modulators (PWM), voltage-to-frequency (V/F) converters, and resistance-to-frequency (R/F) converters. The art of an A/D conversion is well developed. Here, we briefly review some popular architectures of the converters, however, for detailed descriptions the reader should refer to specialized texts, such as [4].

The best known digital code is *binary* (base 2). Binary codes are most familiar in representing integers, i.e., in a natural binary integer code having n bits, the least significant bit (LSB) has a weight of 2^0 (i.e., 1), the next bit has a weight of 2^1 (i.e., 2), and so on up to MSB (most significant bit), which has a weight of 2^{n-1} (i.e., $2^n/2$). The value of a binary number is obtained by adding up the weights of all nonzero bits. When the weighted bits are added up, they form a unique number having any value from 0 to 2^{n-1}. Each additional trailing zero bit, if present, essentially doubles the size of the number.

When converting signals from analog sensors, because full scale is independent of the number of bits of resolution, a more useful coding is fractional binary [4], which is always normalized to full scale. Integer binary can be interpreted as fractional binary if all integer values are divided by 2^n. For example, the MSB has a weight of 1/2 (i.e., $2^{n-1}/2^n = 2^{-1}$), the next bit has a weight of 1/4 (i.e., 2^{-2}), and so forth down to the LSB, which has a weight of $1/2^n$ (i.e., 2^{-n}). When the weighted bits are added up, they form a number with any of 2^n values, from 0 to $(1–2^{-n})$ of full scale. Additional bits simply provide more fine structure without affecting the full-scale range. To illustrate these relationships, Table 5.1 lists 16 permutations of 5-bit's worth of 1's and 0's, with their binary weights, and the equivalent numbers expressed as both decimal and binary integers and fractions.

When all bits are "1" in natural binary, the fractional number value is $1 - 2^{-n}$, or normalized full-scale less 1 LSB (1-1/16=15/16 in the example). Strictly speaking, the number that is represented, written with an "integer point," is 0.1111 (=1 − 0.0001). However, it is almost universal practice to write the code simply as the integer 1111 (i.e., "15") with the fractional nature of the corresponding number understood: "1111" → 1111/(1111 + 1), or 15/16.

For convenience, Table 5.2 lists bit weights in binary for numbers having up to 20 bits. However, the practical range for the vast majority of sensors rarely exceeds 16 bits.

Table 5.1 Integer and fractional binary codes

Decimal fraction	Binary fraction	MSB x1/2	Bit2 x1/4	Bit3 x1/6	Bit4 x1/16	Binary integer	Decimal integer
0	0.0000	0	0	0	0	0000	0
1/16 (LSB)	0.0001	0	0	0	1	0001	1
2/16 = 1/8	0.0010	0	0	1	0	0010	2
3/16 = 1/8 + 1/16	0.0011	0	0	1	1	0011	3
4/16 = 1/4	0.0100	0	1	0	0	0100	4
5/16 = 1/4 + 1/16	0.0101	0	1	0	1	0101	5
6/16 = 1/4 + 1/8	0.0110	0	1	1	0	0110	6
7/16 = 1/4 + 1/8 + 1/16	0.0111	0	1	1	1	0111	7
8/16 = 1/2 (MSB)	0.1000	1	0	0	0	1000	8
9/16 = 1/2 + 1/16	0.1001	1	0	0	1	1001	9
10/16 = 1/2 + 1/8	0.1010	1	0	1	0	1010	10
11/16 = 1/2 + 1/8 + 1/16	0.1011	1	0	1	1	1011	11
12/16 = 1/2 + 1/4	0.1100	1	1	0	0	1100	12
13/16 = 1/2 + 1/4 + 1/16	0.1101	1	1	0	1	1101	13
14/16 = 1/2 + 1/4 + 1/8	0.1110	1	1	1	0	1110	14
15/16 = 1/2 + 1/4 + 1/8 + 1/16	0.1111	1	1	1	1	1111	15

The weight assigned to the LSB is the resolution of numbers having n bits. The dB column represents the logarithm (base 10) of the ratio of the LSB value to unity (full scale), multiplied by 20. Each successive power of 2 represents a change of 6.02 dB [i.e., 20 $\log_{10}(2)$] or "6 dB/octave."

5.5.2 V/F Converters

A voltage-to-frequency (V/F) converter can provide a high-resolution conversion, and such is useful for sensor's special features as a long-term integration (from seconds to years), a digital-to-frequency conversion (together with a D/A converter), a frequency modulation, a voltage isolation, and an arbitrary frequency division and multiplication. The converter accepts an analog output from the sensor, which can be either voltage or current (in latter case, of course, it should be called a current-to-voltage converter). In some cases, a sensor may become a part of an A/D converter as it is illustrated in Sect. 5.6. Here, however, we will discuss only the conversion of voltage to frequency, or, in other words, to a number of square pulses per unit of time. The frequency is a digital format because pulses can be gated (selected for a given interval of time) and then counted, resulting in a binary number. All V/F converters are of the integrating type because the number of pulses per second, or *frequency*, is proportional to the *average* value of the input voltage.

By using a V/F converter, an A/D can be performed in the most simple and economical manner. The time required to convert an analog voltage into a digital number is related to the full-scale frequency of the V/F converter and the required resolution. Generally, the V/F converters are relatively slow, as compared with

Table 5.2 Binary bit weights and resolutions

BIT	2-*n*	1/2*n* fraction	dB	1/2*n* decimal	%	ppm
FS	20	1	0	1.0	100	1,000,000
MSB	2−1	1/2	−6	0.5	50	500,000
2	2−2	1/4	−12	0.25	25	250,000
3	2−3	1/8	−18.1	0.125	12.5	125,000
4	2−4	1/16	−24.1	0.0625	6.2	62,500
5	2−5	1/32	−30.1	0.03125	3.1	31,250
6	2−6	1/64	−36.1	0.015625	1.6	15,625
7	2−7	1/128	−42.1	0.007812	0.8	7,812
8	2−8	1/256	−48.2	0.003906	0.4	3,906
9	2−9	1/512	−54.2	0.001953	0.2	1,953
10	2−10	1/1,024	−60.2	0.0009766	0.1	977
11	2−11	1/2,048	−66.2	0.00048828	0.05	488
12	2−12	1/4,096	−72.2	0.00024414	0.024	244
13	2−13	1/8,192	−78.3	0.00012207	0.012	122
14	2−14	1/16,384	−84.3	0.000061035	0.006	61
15	2−15	1/32,768	−90.3	0.0000305176	0.003	31
16	2−16	1/65,536	−96.3	0.0000152588	0.0015	15
17	2−17	1/131,072	−102.3	0.00000762939	0.0008	7.6
18	2−18	1/262,144	−108.4	0.000003814697	0.0004	3.8
19	2−19	1/524,288	−114.4	0.000001907349	0.0002	1.9
20	2−20	1/1,048,576	−120.4	0.0000009536743	0.0001	0.95

successive approximation devices, however, they are quite appropriate for the vast majority of sensor applications. When acting as an A/D converter, the V/F converter is coupled to a counter, which is clocked with the required sampling rate. For instance, if a full-scale frequency of the converter is 32 kHz, and the counter is clocked 8 times per second, the highest number of pulses, which can be accumulated every counting cycle is 4,000, which approximately corresponds to a resolution of 12 bit (see Table 5.2). By using the same combination of components the V/F converter and the counter, an integrator can be built for the applications, where the stimulus needs to be integrated over a certain time. The counter accumulates pulses over the gated interval rather than as an average number of pulses per counting cycle.

Another useful feature of a V/F converter is that its pulses can be easily transmitted through communication lines. The pulsed signal is much less susceptible to noisy environment than a high-resolution analog signal. In the ideal case, the output frequency f_{out} of the converter is proportional to the input voltage V_{in}:

$$\frac{f_{out}}{f_{FS}} = \frac{V_{in}}{V_{FS}}, \tag{5.22}$$

where f_{FS} and V_{FS} are the full-scale frequency and input voltage, respectively. For a given linear converter, ratio $f_{FS}/V_{FS} = G$ is constant and is called a conversion factor, then

$$f_{out} = GV_{in}. \tag{5.23}$$

There are several known types of V/F converters. The most popular of them are the multivibrator and the charge-balance circuit.

A multivibrator V/F converter employs a free-running square-wave oscillator where charge-discharge currents of a timing capacitor are controlled by the input signal (Fig. 5.24). Input voltage V_{in} is amplified by a differential amplifier (for instance, an instrumentation amplifier) whose output signal controls two voltage-to-current converters (transistors U_1 and U_2). A precision multivibrator alternatively connects timing capacitor C to both current converters. The capacitor is charged for a half of period through transistor U_1 by the current i_a. During the second half of the timing period, it is discharged by the current i_b through transistor U_2. Since currents i_a and i_b are controlled by the input signal, the capacitor charging and discharging slopes vary accordingly, thus changing the oscillating frequency. An apparent advantage of this circuit is its simplicity and potentially very low power consumption, however, its ability to reject high-frequency noise in the input signal is not as good as in the charge-balance architecture.

The charge-balance type of converter employs an analog integrator and a voltage comparator as shown in Fig. 5.25. This circuit has such advantages as high speed, high linearity, and good noise rejection. The circuit is available in an integral form from several manufacturers, for instance, ADVFC32 and AD650 from Analog Devices, LM331 from National Semiconductors. The converter operates as follows. Input voltage V_{in} is applied to an integrator through the input resistor R_{in}. The integrating capacitor is connected as a negative feedback loop to the operational amplifier whose output voltage is compared with a small negative threshold of -0.6 V. The integrator generates a saw-tooth voltage (Fig. 5.27) that at the moment of comparison with the threshold results in a transient at the comparator's output. That transient enables a one-shot generator, which produces a square pulse of a fixed duration t_{os}. A precision current source generates constant current i, which is alternatively applied either to the summing node of the integrator, or to its output. The switch S_1 is controlled by the one-shot pulses. When the current source

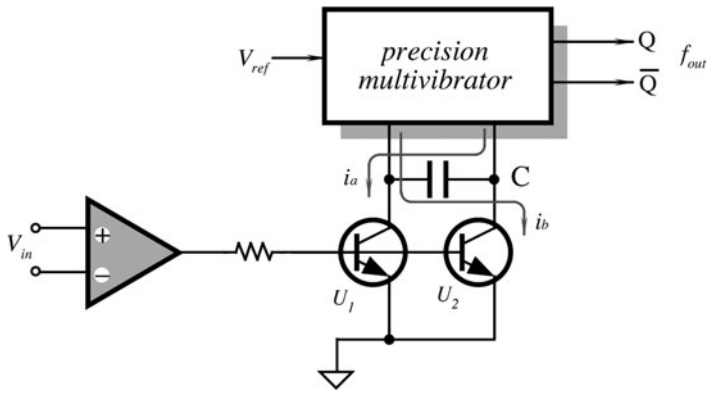

Fig. 5.24 Multivibrator type of a voltage-to-frequency converter

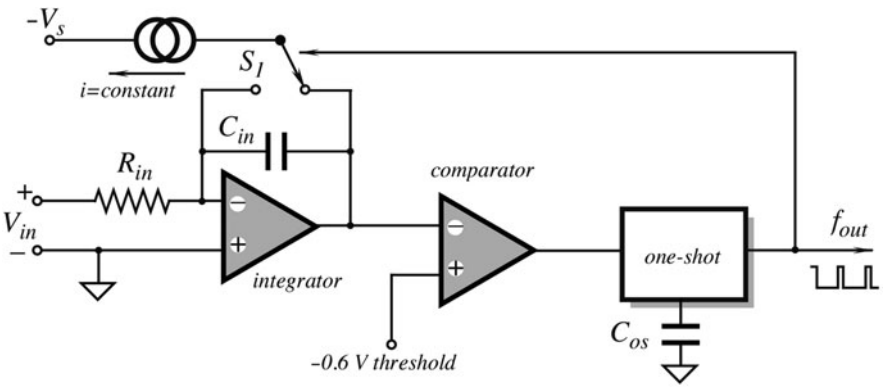

Fig. 5.25 Charge-balance V/F converter

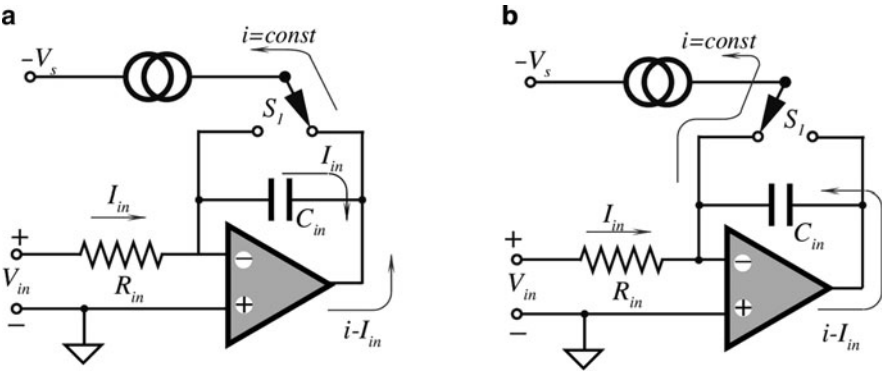

Fig. 5.26 Integrate and deintegrate (reset) phases in a charge-balance converter

is connected to the summing node, it delivers a precisely defined packet of charge $\Delta Q = it_{os}$ to the integrating capacitor. The same summing node also receives an input charge through the resistor R_{in}, thus the net charge is accumulated on the integrating capacitor C_{in}.

When the threshold is reached, the one-shot is triggered and the switch S_1 changes its state to high, thus initiating a reset period (Fig. 5.26b). During the reset period, the current source delivers its current to the summing node of the integrator. The input current charges the integrating capacitor upward. The total voltage between the threshold value and the end of the de-integration is determined by the duration of a one-shot pulse:

$$\Delta V = t_{os}\frac{\mathrm{d}V}{\mathrm{d}t} = t_{os}\frac{i - I_{in}}{C_{in}}. \tag{5.24}$$

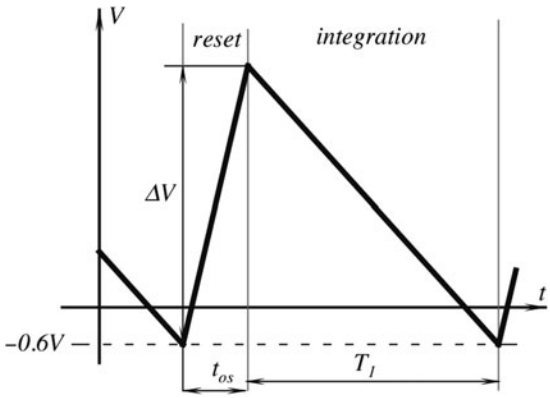

Fig. 5.27 Integrator output in a charge-balance converter

When the output signal of the one-shot circuit goes low, switch S_1 diverts current i to the output terminal of an integrator, which makes no effect on the state of the integrating capacitor C_{in}. That is, the current source sinks a portion of the output current from the operational amplifier. This time is called the integration period (Figs. 5.26a and 5.27). During the integration, the positive input voltage delivers current $I_{in} = V_{in}/R_{in}$ to the capacitor C_{in}. This causes the integrator to ramp down from its positive voltage with the rate proportional to V_{in}. The amount of time required to reach the comparator's threshold is:

$$T_1 = \frac{\Delta V}{dV/dt} = t_{os}\frac{i - I_{in}}{C_{in}}\frac{1}{I_{in}/C_{in}} = t_{os}\frac{i - I_{in}}{I_{in}}. \tag{5.25}$$

It is seen that the capacitor value does not effect duration of the integration period. The output frequency is determined by:

$$f_{out} = \frac{1}{t_{os} + T_1} = \frac{I_{in}}{t_{os} i} = \frac{V_{in}}{R_{in}}\frac{1}{t_{os} i}. \tag{5.26}$$

Therefore, the frequency of one-shot pulses is proportional to the input voltage. It depends also on quality of the integrating resistor, stability of the current generator, and a one-shot circuit. With a careful design, this type of a V/F converter may reach nonlinearity error below 100 ppm and can generate frequencies from 1 Hz to 1 MHz.

A major advantage of the integrating-type converters, such as a charge-balanced V/F converter, is the ability to reject large amounts of additive noise. By integrating of the measurement, noise is reduced or even totally eliminated. Pulses from the converter are accumulated for a gated period T in a counter. Then, the counter behaves like a filter having a transfer function in the form

$$H(f) = \frac{\sin \pi fT}{\pi fT}, \tag{5.27}$$

where f is the frequency of pulses. For low frequencies, value of this transfer function $H(f)$ is close to unity, meaning that the converter and the counter make correct measurements. However, for a frequency $1/T$ the transfer function $H(1/T)$ is zero, meaning that these frequencies are completely rejected. For example, if gating time $T = 16.67$ ms, which corresponds to a frequency of 60 Hz, the power line frequency, which is a source of substantial noise in many sensors, then the 60 Hz noise will be rejected. Moreover, the multiple frequencies (120, 180, 240 Hz, and so on) will also be rejected.

5.5.3 Dual-Slope Converters

A dual-slope converter has been very popular; it was used nearly universally in handheld digital voltmeters and other portable instruments where a fast conversion was not required. This type of converter performs an indirect conversion of the input voltage. First, it converts V_{in} into a function of time, then the time function is converted into a digital number by a pulse counter. The dual slope has the same advantage as the charge-balanced converter; they both reject frequencies $1/T$ corresponding to the integrate timing.

Dual-slope converters are often implemented as a combination of analog components (OPAMs, switches, resistors, and capacitors) and a microcontroller, which handles the functions of timing, control logic, and counting.

5.5.4 Successive Approximation Converter

These converters are widely used in a monolithic form thanks to their high speed (to 1 MHz throughput rates)and high resolution (16 bit and higher). Conversion time is fixed and independent of the input signal. Each conversion is unique, as the internal logic and registers are cleared after each conversion, thus making these A/D converters suitable for the multichannel multiplexing. The converter (Fig. 5.28) consists of a precision voltage comparator, a module comprising shifter registers and a control logic, and a digital-to-analog converter (D/A) that serves as a feedback from the digital outputs to the input analog comparator.

The conversion technique consists of comparing the unknown input, V_{in}, against a precise voltage, V_a, or current generated by a D/A converter. The conversion technique is similar to a weighing process using a balance, with a set of n binary weights (for instance, 1/2 kg, 1/4 kg, 1/8 kg, 1/16 kg, etc. up to total of 1 kg). Before the conversion cycles, all the registers must be cleared and the comparator's output is HIGH. The D/A converter has MSB (1/2 scale) at its inputs and generates an

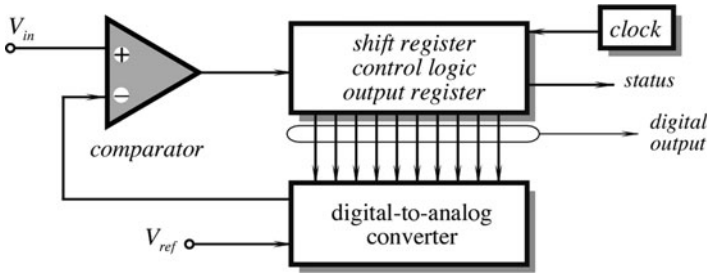

Fig. 5.28 Block-Diagram of successive-approximation A/D converter

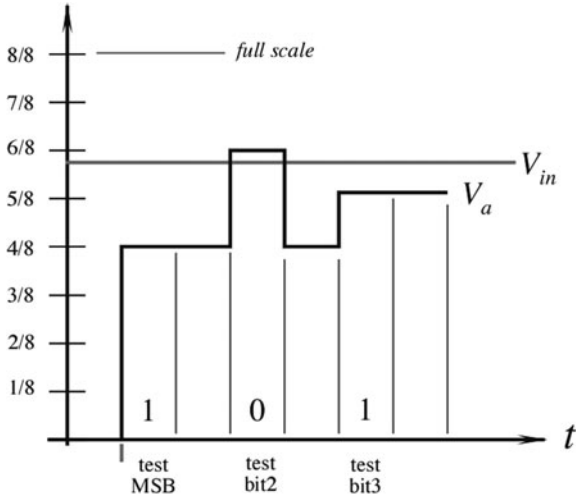

Fig. 5.29 3-bit weighing in successive approximation A/D

appropriate analog voltage, V_a, equal to 1/2 of a full scale input signal. If the input is still greater than the D/A voltage (Fig. 5.29), the comparator remains HIGH, causing "1" at the register's output. Then, the next bit (2/8 = 1/4 of FS) is tried. If the second bit does not add enough weight to exceed the input, the comparator remains HIGH ("1" at the output), and the third bit is tried. However, if the second bit tips the scale too far, the comparator goes LOW resulting in "0" in the register, and the third bit is tried. The process continues in order of descending bit weight until the last bit has been tried. After the completion, the status line indicates the end of conversion and data can be read from the register as a valid number corresponding to the input signal.

To make the conversion valid, the input signal V_{in} must not change until all the bits are tried, otherwise, the digital reading may be erroneous. To avoid any problems with the changing input, a successive approximation converter usually

is supplied with a sample-and-hold (S&H) circuit. This circuit is a short-time analog memory, which samples the input signal and stores it as a dc voltage during an entire conversion cycle.

5.5.5 *Resolution Extension*

In a typical data acquisition system, a monolithic microcontroller often contains an analog-to-digital converter, whose maximum resolution typically is limited to 10 bits. When resolution of a built-in converter is higher, 12 or even 14 bits, the cost may become prohibitively high. In most applications, 10 bits may be not nearly enough for correct representation of a minimum increment of a stimulus (input resolution R_o). There are several ways to resolve this problem. One is to use an amplifier before the A/D converter. For example, an amplifier of gain 4 will effectively increase the input resolution by two bits, say from 10 to 12. Of course, the price to pay is an uncertainty in the amplifier's characteristics. Another method of achieving higher resolution is use a dual-slope A/D converter whose resolution limited only by the available counter rate and the speed response of a comparator.[3] And another method is to use a 10-bit A/D converter (for instance, of a successive approximation type) with a resolution extension circuit. Such a circuit can boost the resolution by several bits, for instance from 10 to 14. A block-diagram of the circuit is shown in Fig. 5.30a. In addition to a conventional 10-bit A/D converter,

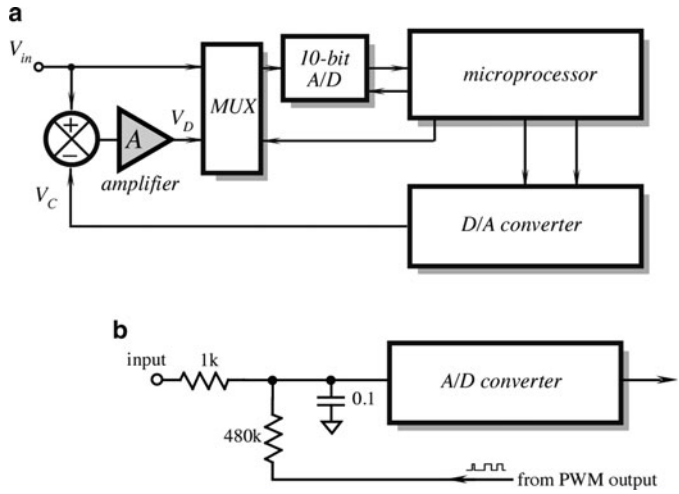

Fig. 5.30 Resolution enhancement circuit with D/A converter (**a**); adding artificial noise to the input signal for oversampling (**b**)

[3] A resolution should not be confused with accuracy.

it includes a digital-to-analog (D/A) converter, a subtraction circuit and an amplifier having gain A. In the ASIC or discrete circuits, a D/A converter may be shared with an A/D part (see Fig. 5.28).

The input signal V_m has a full-scale value E, thus for an 8-bit converter, the initial resolution will be

$$R_o = E/(2^{10} - 1) = E/1023, \tag{5.28}$$

which is expressed in volts per bit. For instance, for a 5 V full scale, the 10-bit resolution is 4.89 mV/bit. Initially, the multiplexer (MUX) connects the input signal to the A/D converter, which produces the output digital value, M, which is expressed in bits. Then, the microprocessor outputs that value to a D/A converter, which produces output analog voltage V_c, which is an approximation of the input signal. This voltage is subtracted from the input signal and amplified by the amplifier to value

$$V_D = (V_m - V_c)A \tag{5.29}$$

The voltage V_D is an amplified error between the actual and digitally represented input signals. For a full scale input signal, the maximum error $(V_m - V_c)$ is equal to a resolution of an A/D converter, therefore, for an 10-bit conversion $V_D = 4.89\text{A mV}$. The multiplexer connects that voltage to the A/D converter which converts V_D to a digital value C:

$$C = \frac{V_D}{R_0} = (V_m - V_c)\frac{A}{R_0}. \tag{5.30}$$

As a result, the microprocessor combines two digital values: M and C, where C represents the high resolution bits. If A=255, then for the 5 V full scale, LSB $\approx$ 19.25 μV, which corresponds to a total resolution of 18 bit. In practice, it is hard to achieve such a high resolution because of the errors originated in the D/A converter, reference voltage, amplifier's drift, noise, etc. Nevertheless, the method is quite efficient when a modest resolution of 12 or 13 bit is deemed to be sufficient.

Another powerful method of a resolution extension is based on the so-called oversampling [18]. The idea works only if the input analog signal is changing between the sampling points. For example, if the A/D conversion steps are at 50, 70, 90 mV, etc. while the input signal is steady 62 mV, the digital number will indicate 70 mV, thus producing a digitization error of 8 mV and no oversampling would make any difference. If the input signal changes with the maximum spectral frequency f_m, according to Nyquist theorem[4] the sampling frequency $f_s > 2f_m$.

[4]A fundamental theorem of the information theory. It is also known as Nyquist-Shannon-Kotelnikov theorem. It states that the minimum sampling must be twice as fast as the highest frequency of the signal.

The oversampling requires a much higher sampling frequency than defined by Nyquist. Specifically, it is based on the formula

$$f_{os} > 2^{2+n} f_m, \tag{5.31}$$

where n is a number of the extension bits. For example, if we have a 10-bit A/D and would like to generate with it a number of 12 bits ($n = 2$), the sampling rate must be at least 16 times higher than f_m. The oversampling allows exchanging a resolution of an A/D conversion for the maximum converted frequency. Thus, this method is useful for converting relatively slow changing signals as compared with the maximum sampling rate of an A/D converter.

As it was said above, the method requires the signal to change between the samplings. If the analog signal does not include natural variations or inherent noise, an artificial noise can be added to the input signal or the A/D reference voltage to jitter signal between the samples. A practical method of adding artificial noise is shown in Fig. 5.30b. The microcontroller generates pulse width modulated (PWM) random pulses that are smoothed by a capacitor and added to the analog input signal. The magnitude of jittering must correspond to at least 0.5 LSB of the original resolution but preferably should be about 2 LSB. After sampling, to get an increased resolution, 2^{2+n} samples from the A/D are added and the result is right-shifted n times. For the above example, 16 sequential 10-bit numbers are added and then right-shifted 2 times, resulting in a 12-bit output number.

5.6 Direct Digitization

Most sensors produce low-level signals. To bring these signals to levels compatible with data processing devices, amplifiers are generally required. Unfortunately, amplifiers and connecting cables and wires may introduce additional errors, add cost to the instrument and increase complexity. Some emerging trends in the sensor-based systems are causing use of the signal conditioning amplifiers to be reevaluated (at least for some transducers) [5]. In particular, many industrial sensor-fed systems are employing digital transmission and processing equipment. These trends point toward direct digitization of sensor outputs, right in the sensor, a difficult task. It is especially true when a sensor-circuit integration on a single chip is considered.

Classical A/D conversion techniques emphasize high-level input ranges. This allows LSB step size to be as large as possible, minimizing offset and noise error. For this reason, a minimum LSB signal is always selected to be at least 100–200 μV. Therefore, a direct connection of many sensors, for instance, RTD temperature transducers or piezoresistive strain gauge s, is unrealistic. In such transducers, a full-scale (FS) output may be limited by several millivolts, meaning that a 10-bit A/D converter must have about 1 μV LSB.

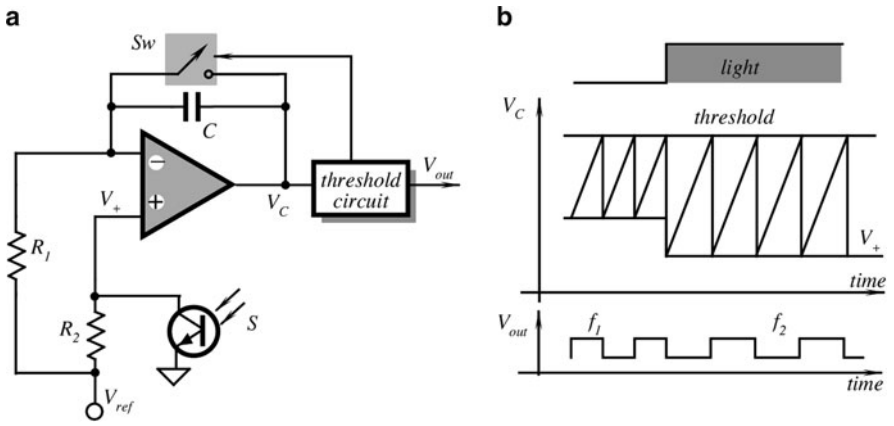

Fig. 5.31 Simplified schematic (**a**) and voltages (**b**) of a light-modulated oscillator

Direct digitization of transducers eliminates a dc gain stage and may yield a better performance without sacrificing accuracy. The main idea behind a direct digitization is to incorporate a sensor into a signal converter, for instance, an A/D converter or an impedance-to-frequency converter. All such converters perform a modulation process and, therefore, are nonlinear devices. Hence, they have some kind of nonlinear circuit, often a threshold comparator. Shifting the threshold level, for instance, may modulate the output signal, which is a desirable effect.

Figure 5.31a shows a simplified circuit diagram of a modulating oscillator. It is comprised of an integrator built with an operational amplifier and a threshold circuit. The voltage across capacitor, C, is an integral of the current whose value is proportional to voltage in the noninverting input of the operational amplifier. When that voltage reaches the threshold, switch SW closes, thus fully discharging the capacitor. The capacitor starts integrating the current again until the cycle repeats. The operating point of the amplifier is defined by the resistor R_2, a phototransistor S, and the reference voltage V_{ref}. A change in light flux, which is incident on the base of the transistor, changes its collector current, thus shifting the operation point. A similar circuit may be used for direct conversion of a resistive transducer, for instance, a thermistor. The circuit can be further modified for the accuracy enhancement, such as for the compensation of the amplifier's offset voltage or bias current, temperature drift, etc.

5.7 Capacitance-to-Voltage Converters

Capacitive sensors are very popular in many applications. Nowadays, micromachining technology allows fabrication of small monolithic capacitive sensors. Capacitive pressure transducers employ a thin silicon diaphragm as a movable plate

of the variable-gap capacitor, which is completed by a metal electrode on the opposing plate. The principle problem in these capacitors is a relatively low capacitance value per unit area (about 2 pF/mm^2) and resulting large die sizes. A typical device offers a zero pressure capacitance on the order of few picofarads, so that a 10-bit resolution requires the detection of capacitive shifts on the order of 15 fF or less (1 femtofarad=10^{-15} F). It is obvious that any external measurement circuit will be totally impractical, as parasitic capacitance of connecting conductors at best can be on the order of 1 pF: too much with respect to the capacitance of the sensor. Therefore, the only way to make such a sensor practical is to build an interface circuit as an integral part of the sensor itself. One quite effective way of designing such a circuit is to use a switched capacitor technique. The technique is based on charge transfer from one capacitor to another by means of solid-state analog switches.

Figure 5.32a shows a simplified circuit diagram of a switched-capacitor converter [6], where variable capacitance C_x and reference capacitance C_r are parts of a symmetrical silicon pressure sensor. Monolithic MOS switches (1–4) are driven by opposite phase clock pulses, ϕ_1 and ϕ_2. When the clocks switch, a charge appears at the common capacitance node. The charge is provided by the constant voltage source, V_{PM}, and is proportional to $(C_x - C_r)$ and, therefore to applied pressure to the sensor. This charge is applied to a charge-to-voltage converter, which includes an operational amplifier, integrating capacitor C_f, and MOS discharge (reset) switch 5. The output signal is variable-amplitude pulses (Fig. 5.32b) which can be transmitted through the communication line and either demodulated to produce linear signal or can be further converted into digital data. So long as the open loop gain of the integrating OPAM is high, the output voltage is insensitive to stray input capacitance, offset voltage, and temperature drift. The minimum detectable signal (noise floor) is determined by the component noise and temperature drifts of the components. The circuit analysis shows that the minimum noise power occurs when

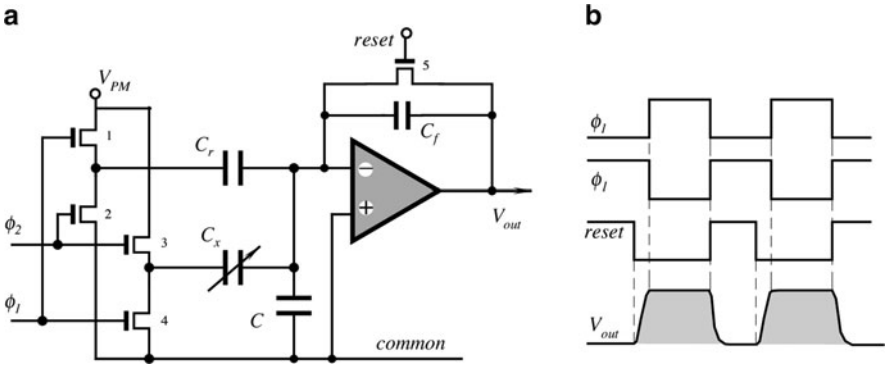

Fig. 5.32 Simplified schematic (**a**) and timing diagrams (**b**) of a differential capacitance-to-voltage converter

the integration capacitor C_f is approximately equal to the frequency compensation capacitor of the OPAM.

When the MOS reset switch goes from the on-state to the off-state, the switching signal injects some charge from the gate of the reset transistor to the input summing node of the OPAM (inverting input). This charge propagated through the gate-to-channel capacitance of the MOS transistor 5. An injection charge results in an offset voltage at the output. This error can be compensated for by a charge-canceling device [7], which can improve the signal-to-noise ratio by two orders of magnitude of the uncompensated charge.

5.8 Integrated Interfaces

A modern trend in the sensor signal conditioning is to integrate in a single silicon chip the amplifiers, multiplexers, A/D converter, and other circuits. Here are two examples of such integration. Figure 5.33 illustrates a signal conditioning circuit ZMD21013 from ZMD (www.zmd.buz). It is optimized for the low-voltage and low-power multiple resistive bridge sensor applications, such as battery-operated consumer or industrial products. This integrated circuit provides programmable amplification and A/D conversion of the sensor signals with up to three resistive bridges or two bridges and one thermocouple, thermopile or any other low-voltage generating sensor. The applied sensor will be switched on only during the sampling

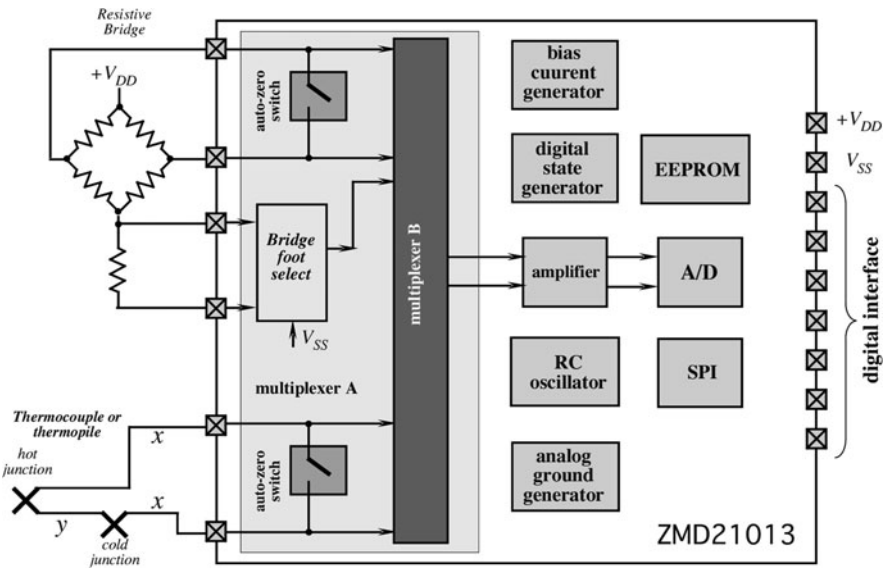

Fig. 5.33 Integrated signal conditioner

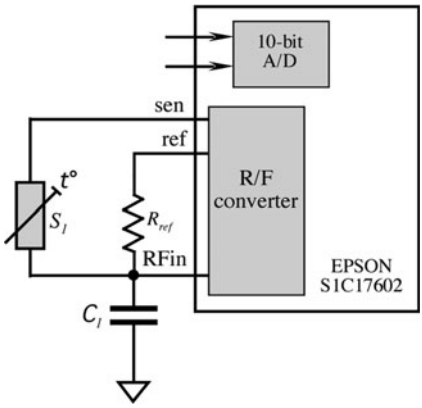

Fig. 5.34 Front end of Epson microcontroller

time, making the circuit suitable for low-power applications. The auto-zero, A/D resolution (10- to 16-bit), sample rate, input range, sensitivity, and measurement mode they all are programmable. This circuit is quite efficient for interfacing a microcontroller with the following resistive and voltage generating sensors: acceleration, pressure, force, flow, and temperature.

Another example of an integrated interface is a 16-bit microcontroller from EPSON having ultralow power consumption. It is specifically adapted for ease of interfacing with various sensors. Figure 5.34 shows that its front end has two types of an A/D converter: a successive approximation 10-bit A/D and an R/F converter with a 24-bit counter. The R/F converter is useful for monitoring temperature and humidity. This microcontroller has various serial interfaces, which enable connection with sensors that are incorporated in healthcare equipment to monitor body temperature, blood pressure, body composition, etc. In addition, this product contains a segment LCD driver for displaying alphanumeric characters and icons. It also contains a built-in MAC (multiply and accumulate) unit and dividing unit enable high-speed processing of the sensor signals. The R/F converter operates with a reference resistor R_{ref} that sets the operating point of the conversion. During the conversion cycle, the capacitor C_1 is charged through a resistive sensor and discharged through R_{ref}, thus value of the capacitor and its stability make no effect on the R/F converter's accuracy.

5.9 Ratiometric Circuits

A powerful method of improving accuracy of a sensor is a *ratiometric* technique, which is one of the most popular ways of signal conditioning. It should be emphasized, however, that the method is useful only if a source of error has a multiplicative nature but not additive. That is, the technique is useless for reduction of, for instance, thermal noise. On the other hand, it is quite potent to solve such

problems as dependence of a sensor's sensitivity to such factors as power supply instability, ambient temperature, humidity, pressure, effects of aging, etc. The technique essentially requires the use of two sensors where one is the acting sensor, which responds to an external stimulus and the other is a compensating sensor, which is either shielded from that stimulus or is insensitive to it. Both sensors must be exposed to all other external effects, which may multiplicatively change their performance. The second sensor, which is often called *reference*, must be subjected to a reference stimulus, which is ultimately stable during the life time of the product. In many practical systems, the reference sensor must not necessarily be exactly similar to the acting sensor, however, its physical properties, which are subject to instabilities should be the same. For example, Fig. 5.35a shows a simple temperature detector where the acting sensor is a negative temperature coefficient (NTC) thermistor R_T. A reference resistor R_o has a value equal to the resistance of the thermistor at some reference temperature, for instance at 25°C. Both are connected via an analog multiplexer to an amplifier with a feedback resistor R. Let us assume that there is a some drift in the sensor value that can be described by a function of time $a(t)$ so that the sensor's resistance becomes $R_T(t) = a(t)R_T$. Property of the resistor R_0 is such that it also changes with the same function, so $R_o(t) = a(t)R_o$. The output signals of the amplifier produced by the sensor and the reference resistor respectively are:

$$V_{\mathrm{N}} = -\frac{ER}{a(t)R_T} = -\frac{ER}{a(t)R_T}, \tag{5.32}$$

$$V_{\mathrm{D}} = -\frac{ER}{a(t)R_o} = -\frac{ER}{a(t)R_o}.$$

It is seen that both voltages are functions of a power supply voltage E and the circuit gain, which is defined by resistor R. They also functions of the drift $a(t)$. The multiplexing switch causes two voltages V_N and V_D to appear sequentially at the

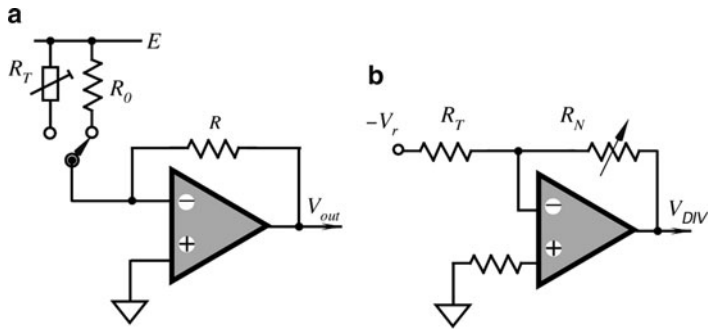

Fig. 5.35 Ratiometric temperature detector (**a**) and analog divider of resistive values (**b**)

amplifier's output. If these voltages are fed into a divider circuit, the resulting signal may be expressed as

$$V_{\mathrm{DIV}} = k\frac{V_{\mathrm{N}}}{V_{\mathrm{D}}} = k\frac{R_{\mathrm{o}}}{R_{\mathrm{T}}}, \tag{5.33}$$

where k is the divider's gain. Therefore, the divider's output signal is subject of neither power supply voltage nor the amplifier gain. It also not subject of the multiplicative drift $a(t)$. The voltage depends only on the sensor and its reference resistor. This is true only if spurious variables, such as function $a(t)$, the power supply or amplifier's gain, do not change rapidly. That is, they should not change appreciably during the multiplexing period. This requirement determines the rate of multiplexing.

A ratiometric technique essentially requires the use of a division. It can be performed by two standard methods: digital and analog. In a digital form, output signals from both the acting and the reference sensors are multiplexed and converted into binary codes in an analog-to-digital (A/D) converter. Subsequently, a computer or a microprocessor performs an operation of a division. In an analog form, a divider may be part of a signal conditioner or the interface circuit. A "divider" [Fig. 5.36a] produces an output voltage or current proportional to the ratio of two input voltages or currents:

$$V_{DIV} = k\frac{V_N}{V_D}, \tag{5.34}$$

where the numerator is denoted as V_N, the denominator V_D and k is equal to the output voltage, when $V_N = V_D$. The operating ranges of the variables (quadrants of operation) is defined by the polarity and magnitude ranges of the numerator and denominator inputs, and of the output. For instance, if V_N and V_D are both either positive or negative, the divider is of a 1-quadrant type. If the numerator is bipolar, the divider is 2-quadrant. Generally, the denominator is restricted to a single polarity,

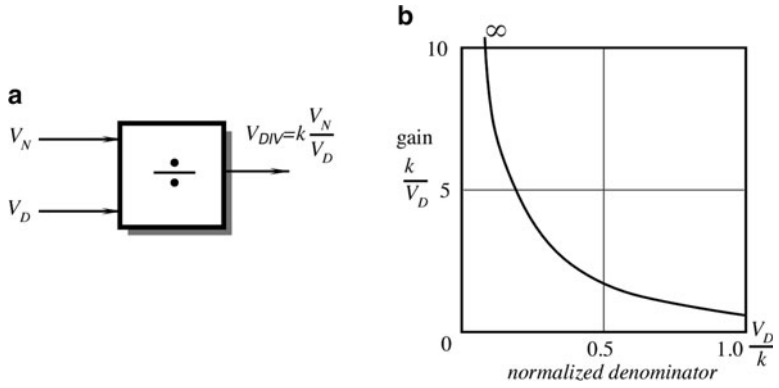

Fig. 5.36 Symbol of a divider (**a**) and gain of a divider as function of a denominator (**b**)

since the transition from one polarity to another would require the denominator to pass through zero, which would call for an infinite output (unless the numerator is also zero). In practice, the denominator is a signal from a reference sensor, which usually is of a relatively constant value.

Division has long been the most difficult of the four arithmetic functions to implement with analog circuits. This difficulty stems primarily from the nature of division: the magnitude of a ratio becomes quite large, approaching infinity, for a denominator that is approaching zero (and a nonzero numerator). Thus, an ideal divider must have a potentially infinite gain and infinite dynamic range. For a real divider, both of these factors are limited by the magnification of drift and noise at low values of V_D. That is, the gain of a divider for a numerator is inversely dependent on the value of the denominator [Fig. 5.36b]. Thus, the overall error is the net effect of several factors, such as gain dependence of denominator, numerator, and denominator input errors, like offsets, noise and drift (which must be much smaller than the smallest values of input signals). Besides, the output of the divider must be constant for constant ratios of numerator and denominator, independent of their magnitudes. For example, $10/10 = 0.01/0.01 = 1$ and $1/10 = 0.001/0.01 = 0.1$. In practice, some simple division circuits are used quite extensively. An example is an amplifier of Fig. 5.35b whose output signal is function of the resistor ratio (note that the reference voltage V_r is negative):

$$V_{DIV} = V_r \frac{R_N}{R_T}, \tag{5.35}$$

The most popular and efficient ratiometric circuits are based on Wheatstone bridge designs which are covered below.

5.10 Differential Circuits

Beside multiplicative interferences, the additive interferences are very common and pose a serious problem for low-level output signals. Consider for example a pyroelectric sensor [Fig. 14.22a] where a heat flow sensitive ceramic plate is supported inside a metal can. Since a pyroelectric is also a piezoelectric, besides heat flow the sensor is susceptible to mechanical stress interferences. Even a slight vibration will generate a spurious piezoelectric signal that may be several orders of magnitude higher than a pyroelectric current. The solution is to fabricate a sensor with dual electrodes deposited on the same ceramic substrate as shown in Fig. 14.22b. This essentially creates two identical sensors on the same ceramic plate. Both sensors respond to all stimuli nearly identically. Since they are oppositely connected and assuming that V_{pyro} and V_{piezo} from one sensor are, respectively, equal to those of the other sensor, the resulting output voltage is essentially zero:

$$V_{out} = (V_{pyro_1} + V_{piezo_1}) - (V_{pyro_2} + V_{piezo_2}) = 0 \tag{5.36}$$

If one of the sensors is blocked from receiving thermal radiation ($V_{pyro_2} = 0$), then $V_{out} = V_{pyro_1}$. In other words, thanks to subtraction ($V_{piezo_1} = V_{piezo_2}$) are cancelling each other), the combined sensor now is insensitive to a piezoelectric effect. A differential method where a sensor is fabricated is a symmetrical form and connected to a symmetrical interface circuit (e.g., differential amplifier) so that one signal is subtracted from another and is a very powerful way of noise and drift reductions. Yet, this method is effective only if a dual sensor is fully symmetrical. An asymmetry will produce a proportional loss of noise cancelation. For example, if asymmetry is 5%, the noise will be cancelled by no more than 95%.

5.11 Bridge Circuits

5.11.1 General Concept

The Wheatstone bridge circuits are popular and very effective implementations of the ratiometric technique (division) technique on a sensor level. A basic circuit is shown in Fig. 5.37. Impedances **Z** may be either active or reactive, that is they may be either simple resistances, like in the piezoresistive gauges, or capacitors, or inductors, or combinations of the above. For a pure resistor, the impedance is R, for an ideal capacitor, the magnitude of its impedance is equal to $1/(2\pi fC)$ and for an inductor, it is $2\pi fL$, where f is the frequency of the current passing through the element. The bridge output voltage is represented by

$$V_{out} = \left(\frac{\mathbf{Z_1}}{\mathbf{Z_1} + \mathbf{Z_2}} - \frac{\mathbf{Z_3}}{\mathbf{Z_3} + \mathbf{Z_4}} \right) V_{ref}, \tag{5.37}$$

The bridge is considered to be in a balanced state when the following condition is met:

$$\frac{\mathbf{Z_1}}{\mathbf{Z_2}} = \frac{\mathbf{Z_3}}{\mathbf{Z_4}}. \tag{5.38}$$

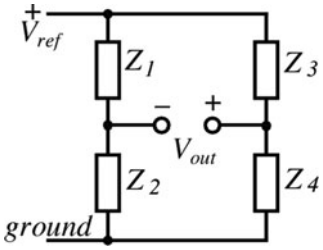

Fig. 5.37 General circuit of Wheatstone bridge

Under the balanced condition, the output voltage is zero. When at least one impedance in the bridge changes, the bridge becomes imbalanced and the output voltage goes either in a positive or negative direction, depending on the direction of the impedance change. To determine the bridge sensitivity with respect to each impedance partial derivatives may be obtained from (5.37):

$$\begin{aligned}
\frac{\partial V_{out}}{\partial \mathbf{Z_1}} &= \frac{\mathbf{Z}_2}{(\mathbf{Z_1}+\mathbf{Z_2})^2} V_{ref} \\
\frac{\partial V_{out}}{\partial \mathbf{Z_2}} &= -\frac{\mathbf{Z}_1}{(\mathbf{Z_1}+\mathbf{Z_2})^2} V_{ref} \\
\frac{\partial V_{out}}{\partial \mathbf{Z_3}} &= -\frac{\mathbf{Z}_4}{(\mathbf{Z_3}+\mathbf{Z_4})^2} V_{ref} \\
\frac{\partial V_{out}}{\partial \mathbf{Z_4}} &= \frac{\mathbf{Z}_3}{(\mathbf{Z_3}+\mathbf{Z_4})^2} V_{ref}
\end{aligned} \tag{5.39}$$

By summing these equations, we obtain the bridge sensitivity:

$$\frac{\delta V_{out}}{V_{ref}} = \frac{\mathbf{Z}_2 \delta \mathbf{Z_1} - \mathbf{Z_1} \delta Z_2}{(\mathbf{Z_1}+\mathbf{Z_2})^2} - \frac{\mathbf{Z}_4 \delta \mathbf{Z_3} - \mathbf{Z_3} \delta \mathbf{Z_4}}{(\mathbf{Z_3}+\mathbf{Z_4})^2}, \tag{5.40}$$

A closer examination of (5.40) shows that only the adjacent pairs of the impedances (i.e., $\mathbf{Z_1}$ and $\mathbf{Z_2}$, $\mathbf{Z_3}$, and $\mathbf{Z_4}$) have to be identical in order to achieve the ratiometric compensation (such as the temperature stability, drift, etc.). It should be noted that impedances in the balanced bridge do not have to be equal, as long as a balance of the ratio (5.38) is satisfied. In many practical circuits, only one impedance is used as a sensor, thus for $\mathbf{Z}_1$ as a sensor, the bridge sensitivity becomes

$$\frac{\delta V_{\text{out}}}{V_{\text{ref}}} = \frac{\delta \mathbf{Z_1}}{4\mathbf{Z_1}}. \tag{5.41}$$

The resistive bridge circuits are commonly used with strain gauges, piezoresistive pressure transducers, thermistor thermometers, hygristors, and other sensors when immunity against environmental factors is required. Similar arrangements are used with the capacitive and magnetic sensors for measuring force, displacement, moisture, etc.

5.11.2 Disbalanced Bridge

A basic Wheatstone bridge circuit [Fig. 5.38a] generally operates with a disbalanced bridge. This is called the *deflection* method of measurement. It is based on a detecting the voltage across the bridge diagonal. The bridge output voltage is

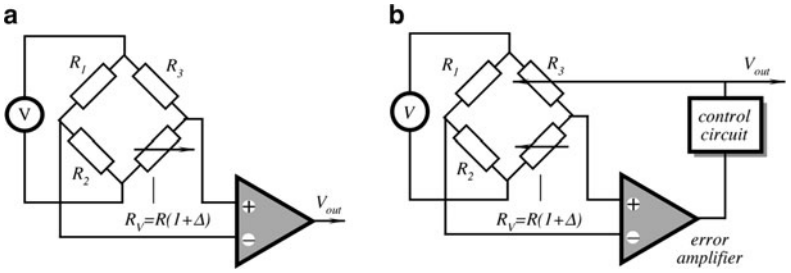

Fig. 5.38 Two methods of using a bridge circuit: Disbalanced bridge (**a**) and Null-balanced bridge with a feedback control (**b**)

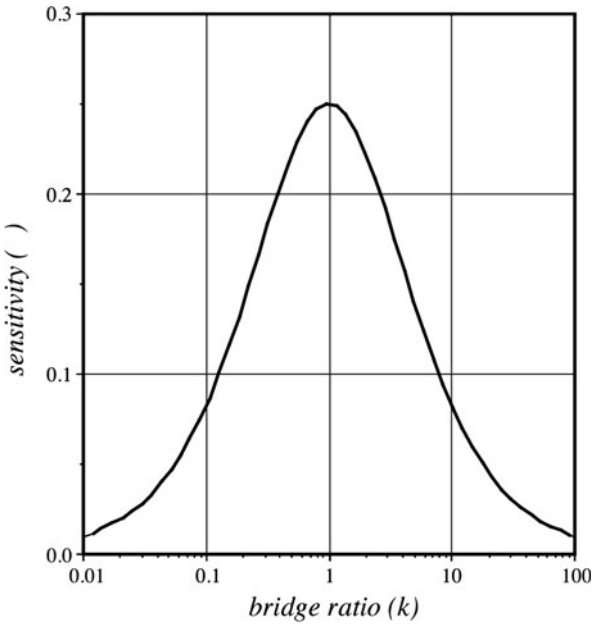

Fig. 5.39 Sensitivity of a disbalanced bridge as function of the impedance ratio

a nonlinear function of a disbalance Δ, where the sensor's resistance $R_v = R(1 + \Delta)$. However, for a small change ($\Delta < 0.05$), which often is the case, the bridge output may be considered quasi-linear. The bridge maximum sensitivity is obtained when $R_1 = R_2$ and $R_3 = R$. When $R_1 >> R_2$ or $R_2 >> R_1$, the bridge output voltage is decreased. Assuming that $k = R_1/R_2$, the bridge sensitivity may be expressed as:

$$\alpha = \frac{V}{R} \frac{k}{(k+1)^2} \tag{5.42}$$

A normalized graph calculated according to this equation is shown in Fig. 5.39. It indicates that the maximum sensitivity is achieved at $k = 1$, however, the sensitivity drops relatively little for the range where $0.5 < k < 2$. If the bridge is fed by a current source, rather by a voltage source, its output voltage for small Δ and a single variable component is represented by

$$V_{out} = i\frac{k\Delta}{2(k+1)}, \tag{5.43}$$

where i is the excitation current.

5.11.3 Null-Balanced Bridge

Another method of using a bridge circuit is called a *null-balance*. The method overcomes the limitation of small changes (Δ) in the bridge arm to achieve a good linearity. The null-balance essentially requires that the bridge is *always* maintained at the balanced state. To satisfy the requirement for a bridge balance (5.37) another arm of the bridge should vary along with the sensing arm. Figure 5.38b illustrates this concept. A control circuit modifies the value of R_3 on a command from the error amplifier. The sensor's output voltage may be obtained from the control signal of the balancing arm R_3. For example, both R_v and R_3 may be photoresistors. The R_3-photoresistor could be interfaced with a light emitting diode (LED), which is controlled by the error amplifier. Current through the LED becomes a measure of resistance R_v, and, subsequently, of the light intensity detected by the sensor.

5.11.4 Bridge Amplifiers

The bridge amplifiers for resistive sensors are probably the most frequently used sensor interface circuits. They may be of various configurations, depending on the required bridge grounding and availability of either grounded or floating reference voltages. Figure 5.40a shows the so-called active bridge, where a variable resistor (the sensor) is floating, i.e., isolated from ground, and is connected into a feedback of the OPAM. If a resistive sensor's transfer function can be modeled by a first order function:

$$R_x \approx R_o(1+\alpha), \tag{5.44}$$

where α is the input stimulus, a transfer function of this circuit is

$$V_{out} = \frac{V}{2} - \frac{1}{2}\alpha V. \tag{5.45}$$

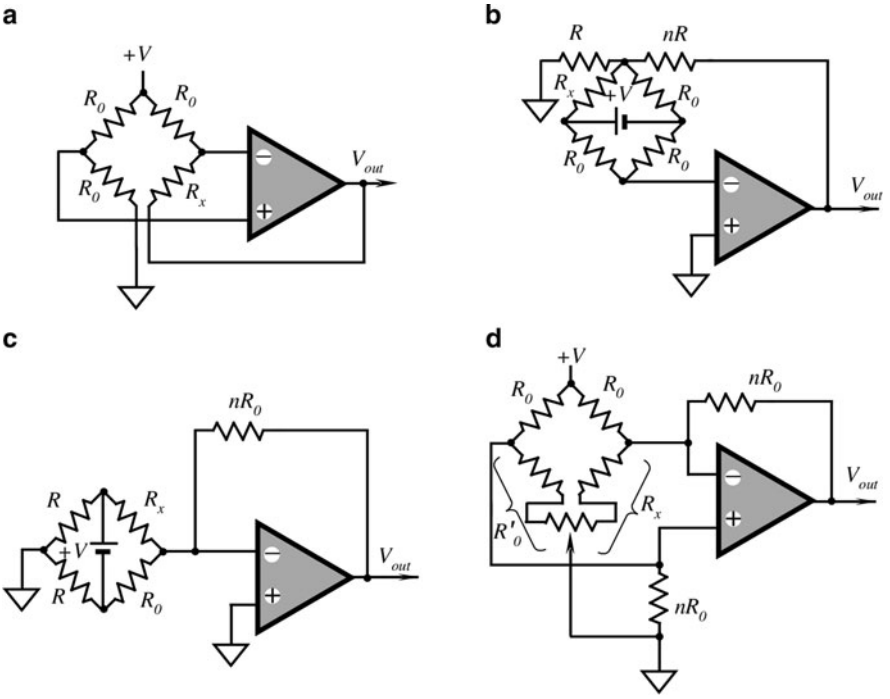

Fig. 5.40 Connection of operational amplifiers to resistive bridge circuits (disbalanced mode)

A circuit with a floating bridge and floating reference voltage source V is shown in Fig. 5.40b. This circuit may provide gain which is determined by a feedback resistor whose value is nR_o:

$$V_{out} = \frac{V}{2} + (1+n)\alpha\frac{V}{4}\frac{1}{1+\frac{\alpha}{2}} \approx \frac{V}{2}\left(1 + \frac{(1+n)\alpha}{2}\right). \tag{5.46}$$

A bridge with the asymmetrical resistors ($R \neq R_o$) may be used with the circuit shown in Fig. 5.40c. It requires a floating reference voltage source V:

$$V_{\text{out}} = \frac{V}{2} + n\alpha\frac{V}{4}\frac{1}{1+\frac{\alpha}{2}} \approx \frac{V}{2}\left(1 + \frac{n\alpha}{2}\right) \tag{5.47}$$

When a resistive sensor is grounded and a gain from the interface circuit is desirable, a schematic shown in Fig. 5.40d may be employed. Its transfer function is determined from

$$V_{out} = \frac{V}{2} - \frac{n}{2}\frac{V}{1+\frac{1}{2n}}\frac{\alpha}{1+\alpha} \approx 0.5V\left(1 - \frac{n}{1+\frac{1}{2n}}\alpha\right) \tag{5.48}$$

Note that the circuit may contain a balancing potentiometer whose resistance sectors should be included into the corresponding arms of the bridge. The potentiometer is used to adjust the bridge component tolerances or offset the bridge balance by some fixed bias. When the bridge is perfectly balanced, its output voltage V_{out} is equal to a half of the bridge excitation voltage $+V$. To better utilize the operational amplifier open loop gain, the value of n should not exceed 100.

5.12 Data Transmission

Signal from a sensor may be transmitted to a receiving end of the system either in a digital format or analog. In most cases, a digital format essentially requires use of an analog-to-digital converter at the sensor's site. Transmission in a digital format has several advantages, the most important of which is noise immunity. Since transmission of digital information is beyond the scope of this book we will not discuss it further. In many cases, however, digital transmission can not be done for several reasons. Then, the sensor's output signal is transmitted to the receiving site in an analog form. Depending on connection, the transmission methods can be divided into a 2, 4, and 6-wire methods.

5.12.1 Two-Wire Transmission

Two-wire analog transmitters are used to couple sensors to control and monitoring devices in the process industry [10]. When, for example, a temperature measurement is taken within a process, a 2-wire transmitter relays that measurement to the control room or interfaces the analog signal directly to a process controller. Two wires can be used to transmit either voltage or current, however, current was accepted as an industry standard. The current carried by the wires varies in the range from 4 to 20 mA, which represents the entire span of an input stimuli. Zero stimulus corresponds to 4 mA while the maximum is at 20 mA. There are two advantages of using current rather than voltage as it is illustrated in Fig. 5.41. Two wires link the controller site to the sensor site. On the sensor site, there is a sensor which is connected to the so-called *two-wire transmitter*. The transmitter may be a voltage-to-current converter. That is, it converts the sensor signal into a variable current. On the controller site, there is a voltage source that can deliver current up to 20 mA. The two wires form a current loop, which at the sensor's side has the sensor and a transmitter, while at the controller side it has a load resistor and a power supply, which are connected in series. When the sensor signal varies, the transmitter's output resistance varies accordingly, thus

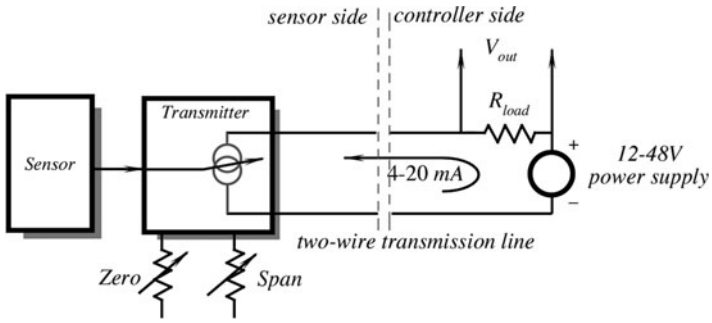

Fig. 5.41 Two-wire 20 mA analog data transmission

modulating the current in the range between 4 and 20 mA. The same current, which carries information, is also used by the transmitter and the sensor to provide their operating power. Obviously, even for the lowest output signal, which produces 4 mA current, the produced voltage must be sufficient to power the transmitting side of the loop. The loop current causes a voltage drop across the load resistor at the controller side. This voltage is a received signal, which is suitable for further processing by the electronic circuits. An advantage of the two-wire method is that the transmitting current is independent of the connecting wires resistance (as long as they do not change) and thus of the transmission line length, obviously, within the limits.

5.12.2 Four-Wire Sensing

Sometimes, it is desirable to connect a resistive sensor to a remotely located interface circuit. When such a sensor has a relatively low resistance (for instance, it is normal for the piezoresistors or RTDs to have resistances on the order of 100 Ω), the connecting wire resistances pose a serious problem since they alter the excitation voltage across the sensor. The problem can be solved by using the so-called 4-wire method (Fig. 5.42a). It allows measuring the resistance of a remote resistor without accounting for resistances of the connecting conductors. A sensor, which is the subject of measurement, is connected to the interface circuit through four rather than two wires. Two wires are connected to a current source and two others to a voltmeter or amplifier. A constant current source (current pump) has a very high output resistance, therefore the current, which it pushes through the loop is almost independent of any resistances r in that loop. An input impedance of a voltmeter or amplifier is very high, hence no current is diverted from the current loop to the voltmeter. A voltage drop across the resistor R_x is

$$V_x = R_x i_o, \tag{5.49}$$

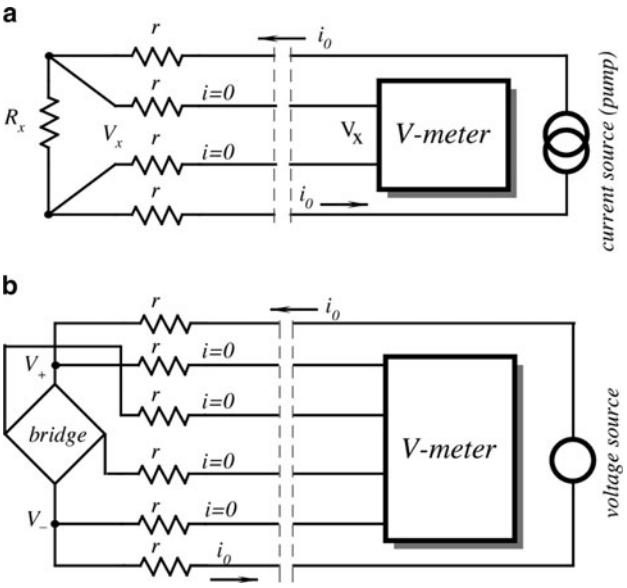

Fig. 5.42 Remote measurements of resistances four-wire method (**a**); six-wire measurement of a bridge (**b**)

which is independent of any resistances r of the connecting wires. The 4-wire method is a very powerful means of measuring resistances of remote detectors and is used in industry and science quite extensively.

5.12.3 Six-Wire Sensing

When a Wheatstone bridge circuit is remotely located, voltage across the bridge plays an important role in the bridge temperature stability. That voltage often should be either measured or controlled. Long transmitting wires may introduce unacceptably high resistance in series with the bridge excitation voltage, which interferes with the temperature compensation. The problem may be solved by providing two additional wires to feed the bridge with voltage and to dedicate two wires to measuring the voltage across the bridge (Fig. 5.42b). The actual excitation voltage across the bridge and the bridge differential output voltage are measured by a high-input impedance voltmeter with negligibly small input currents. Thus, the accurate bridge voltages are available at the data processing site without being affected by long transmission lines.

5.13 Noise in Sensors and Circuits

Noise in sensors and circuits may present a substantial source of errors and should be seriously considered. “Like diseases, noise is never eliminated, just prevented, cured, or endured, depending on its nature, seriousness, and the cost/difficulty of treating” [11]. There are two basic classifications of noise for a given circuit: they are inherent noise, which is noise arising within the circuit, and interference (transmitted) noise, which is noise picked up from outside the circuit.

Any sensor, no matter how well it was designed, never produces an electric signal that is an ideal representation of the input stimulus. Often, it is a matter of judgment to define the goodness of the signal. The criteria for this are based on the specific requirements to accuracy and reliability. Distortions of the output signal can be either systematic or stochastic. The former are related to the sensor’s transfer function, its linearity, dynamic characteristics, etc. They all are the result of the sensor’s design, manufacturing tolerances, material quality, and calibration. During a reasonably short time, these factors either do not change or drift relatively slowly. They can be well-defined, characterized, and specified (see Chap. 2). In many applications, such a determination may be used as a factor in the error budget and can be accounted for. Stochastic disturbances, on the other hand, often are irregular, unpredictable to some degree and may change rapidly. Generally, they are termed noise, regardless of their nature and statistical properties. It should be noted that word noise, in association with audio equipment noise, is often mistaken for an irregular, somewhat fast changing signal. We use this word in a much broader sense for all disturbances, either in stimuli, environment, or in the components of sensors and circuits from dc to the upper operating frequencies.

5.13.1 Inherent Noise

A signal, which is amplified and converted from a sensor into a digital form, should be regarded not just by its magnitude and spectral characteristics, but also in terms of a digital resolution. When a conversion system employs an increased digital resolution, the value of the least-significant bit (LSB) decreases. For example, the LSB of a 10-bit system with a 5 V full scale is about 5 mV, the LSB of 16 bits is 77 μV. This by itself poses a significant problem. It makes no sense to employ, say a 16-bit resolution system, if a combined noise is, for example, 300 μV. In a real world, the situation is usually much worse. There are almost no sensors that are capable of producing a 5 V full-scale output signals. Most of them require an amplification. For instance, if a sensor produces a full-scale output of 5 mV, at a 16-bit conversion it would correspond to a LSB of 77 nV, an extremely small signal which makes amplification an enormous task by itself. Whenever a high resolution of a conversion is required, all sources of noise must be seriously considered. In the circuits, noise can be produced by the monolithic amplifiers

and other components, which are required for the feedback, biasing, bandwidth limiting, etc.

Input offset voltages and bias currents may drift. In dc circuits, they are indistinguishable from low magnitude signals produced by a sensor. These drifts are usually slow (within a bandwidth of tenths and hundredths of a Hz), therefore they are often called ultralow frequency noise. They are equivalent to randomly (or predictable, say with temperature) changing voltage and current offsets and biases. To distinguish them from the higher frequency noise, the equivalent circuit (Fig. 5.3) contains two additional generators. One is a voltage offset generator e_0 and the other is a current bias generator i_0. The noise signals (voltage and current) result from physical mechanisms within the resistors and semiconductors that are used to fabricate the circuits. There are several sources of noise whose combined effect is represented by the noise voltage and current generators.

One cause for noise is a discrete nature of electric current because current flow is made up of moving charges, and each charge carrier transports a definite value of charge (charge of an electron is 1.6×10^{-19} C). At the atomic level, current flow is very erratic. The motion of the current carriers resembles popcorn popping. This was chosen as a good analogy for current flow and has nothing to do with the "popcorn noise," which we will discuss below. As popcorn, the electron movement may be described in statistical terms. Therefore, one never can be sure about very minute details of current flow. The movement of carriers is temperature related and noise power, which in turn, is also temperature related. In a resistor, these thermal motions cause Johnson noise to result [12]. The mean-square value of noise voltage (which is representative of noise power) can be calculated from

$$\bar{e}_n^2 = 4kTR\Delta f \ [\mathrm{V}^2/\mathrm{Hz}], \tag{5.50}$$

where $k = 1.38 \times 10^{-23}$ J/K (Boltzmann constant), T is temperature in K, R is the resistance in Ω, and Δf is the bandwidth over which the measurement is made, in Hz.

For practical purposes, noise density per $\sqrt{Hz}$ generated by a resistor at room temperature may be estimated from a simplified formula $\overline{e_n} \approx 0.13\sqrt{\mathrm{Hz}}$ in $\mathrm{nV}\sqrt{\mathrm{Hz}}$. For example, if noise bandwidth is 100 Hz and the resistance of concern is 10 MΩ (10^7 Ω), the average noise voltage is estimated as $\overline{e_n} \approx 0.13\sqrt{10^7}\sqrt{100} = 4,111$ nV ≈ 4 μV.

Even a simple resistor is a source of noise. It behaves as a perpetual generator of electric signal. Naturally, relatively small resistors generate extremely small noise, however, in some sensors Johnson noise must be taken into account. For instance, a pyroelectric detector uses a bias resistor on the order of 50 GΩ. If a sensor is used at room temperature within a bandwidth of 100 Hz, one may expect the average noise voltage across the resistor to be on the order of 0.3 mV, a pretty high value. To keep noise at bay, bandwidths of the interface circuits must be maintained small, just wide enough to pass the minimum required signal. It should be noted that noise voltage is proportional to square root of the bandwidth. It implies that if we reduce the bandwidth 100 times, noise voltage will be reduced by a factor of 10. Johnson

noise magnitude is constant over a broad range of frequencies. Hence, it is often called white noise because of the similarity to white light, which is composed of all the frequencies in visible spectrum.

Another type of noise results because of dc current flow in semiconductors. It is called shot noise, the name suggested by Schottky not in association with his own name but rather because this noise sounded like "a hail of shot striking the target" (nevertheless, shot noise is often called Schottky noise). Shot noise is also white noise. Its value becomes higher with the increase in the bias current. This is the reason why in FET and CMOS semiconductors the current noise is quite small. For a bias current of 50 pA, it is equal to about $4\,\text{fA}/\sqrt{Hz}$, an extremely small current that is equivalent to movement of about 6,000 electrons per second. A convenient equation for shot noise is

$$i_{\text{sn}} = 5.7 \times 10^{-4}\sqrt{I\Delta f}, \tag{5.51}$$

where I is a semiconductor junction current in picoamperes and Δf is a bandwidth of interest in Hz.

An additional ac noise mechanism exists at low frequencies (Fig. 5.43). Both the noise voltage and noise current sources have a spectral density roughly proportional to $1/f$, which is called the pink noise, because of the higher noise contents at lower frequencies (lower frequencies are also at red side of the visible spectrum). This $1/f$ noise occurs in all conductive materials, therefore it is also associated with resistors. At extremely low frequencies it is impossible to separate the $1/f$ noise from dc drift effects. The $1/f$ noise is sometimes called a flicker noise. Mostly it is pronounced at frequencies below 100 Hz, where many sensors operate. It may dominate Johnson and Schottky noise and becomes a chief source of errors at these frequencies. The magnitude of pink noise depends on current passing through the resistive or semiconductive material. Nowadays progress in semiconductor technology resulted in significant reduction of $1/f$ noise in semiconductors, however, when designing a circuit, it is a good engineering practice to use metal film or wirewound resistors in sensors and the front stages of interface circuits wherever significant currents flow through the resistor and low noise at low frequencies is a definite requirement.

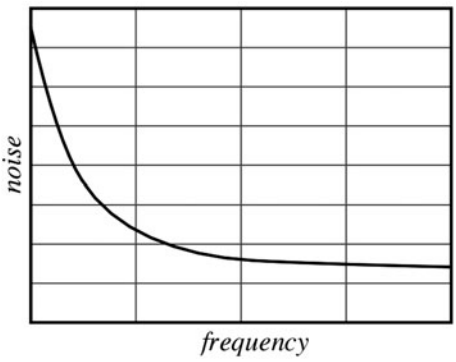

Fig. 5.43 Spectral distribution of $1/f$ "pink" noise

A peculiar ac noise mechanism is sometimes seen on the screen of an oscilloscope when observing the output of an operational amplifier, a principal building block of many sensor interface circuits. It looks like a digital signals transmitted from outer space; noise has a shape of square pulses having variable duration of many milliseconds. This abrupt type of noise is called popcorn noise because of the sound it makes coming from a loudspeaker. Popcorn noise is caused by defects that are dependent on the integrated circuits manufacturing techniques. Thanks to advances fabricating technologies, this type of noise is drastically reduced in modern semiconductor devices.

A combined noise from all voltage and current sources is given by sum of squares of individual noise voltages:

$$e_E = \sqrt{e_{n1}^2 + e_{n2}^2 + \cdots + (R_1 i_{n1})^2 + (R_1 i_{n2})^2 + \cdots}. \tag{5.52}$$

A combined random noise may be presented by its root mean square (r.m.s.) value, that is

$$E_{rms} = \sqrt{\frac{1}{T}\int_0^T e^2 dt}, \tag{5.53}$$

where T is time of observation, e is noise voltage and t is time.

Also, noise may be characterized in terms of the peak values, which are the differences between the largest positive and negative peak excursions observed during an arbitrary interval. For some applications, in which peak-to-peak (*p-p*) noise may limit the overall performance (in a threshold-type devices), *p-p* measurement may be essential. Yet, due to a generally Gaussian distribution of noise signal, *p-p* magnitude is very difficult to measure in practice. Because *r.m.s.* values are so much easier to measure repeatedly, and they are the most usual form for presenting noise data noncontroversially, the Table 5.3 should be useful for estimating the probabilities of exceeding various peak values given by the *r.m.s.* values. The casually observed *p-p* noise varies between 3 × r.m.s. and 8 × *r.m.s.*, depending on the patience of observer and amount of data available.

Table 5.3 Peak-to-peak value vs. *r.m.s.* (for Gaussian distribution)

Nominal p-p voltage	% of time that noise will exceed nominal p-p value
2 × r.m.s.	32.0%
3 × r.m.s.	13.0%
4 × r.m.s.	4.6%
5 × r.m.s.	1.2%
6 × r.m.s.	0.27%
7 × r.m.s.	0.046%
8 × r.m.s.	0.006%

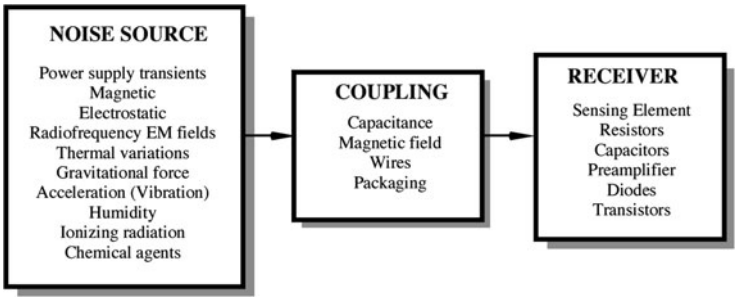

Fig. 5.44 Sources and coupling of transmitted noise

5.13.2 Transmitted Noise

A large portion of environmental stability is attributed to immunity of a sensor and an interface circuit to noise, which is originated in external sources. Figure 5.44 shows a diagram of the transmitted noise propagation. Noise comes from a source which often cannot be identified. Examples of the sources are: voltage surges in power lines, lightnings, change in ambient temperature, sun activity, etc. These interferences propagate toward the sensor and the interface circuit, and to present a problem eventually must appear at the output. However, before that, they somehow must affect the sensing element inside the sensor, its output terminals or the electronic components in the circuit. Both the sensor and circuit act as receivers of the interferences.

There can be several classifications of transmitted noise, depending on how it affects the output signal, how it enters the sensor or circuit, etc. With respect to its relation to the output signals, noise can be either additive or multiplicative.

Additive noise e_n is added to the useful signal V_s and mixed with it as a fully independent voltage (or current)

$$V_{out} = V_s + e_n. \tag{5.54}$$

An example of such a disturbance is depicted in Fig. 5.45b. It can be seen that the noise magnitude does not change when the actual (useful) signal changes. As long as the sensor and interface electronics can be considered linear, the additive noise magnitude is totally independent of the signal magnitude and, if the signal is equal to zero, the output noise still will be present.

Multiplicative noise affects the sensor's transfer function or the circuit's nonlinear components in such a manner as V_s signal's value becomes altered or *modulated* by the noise:

$$V_{out} = [1 + N(t)]V_s, \tag{5.55}$$

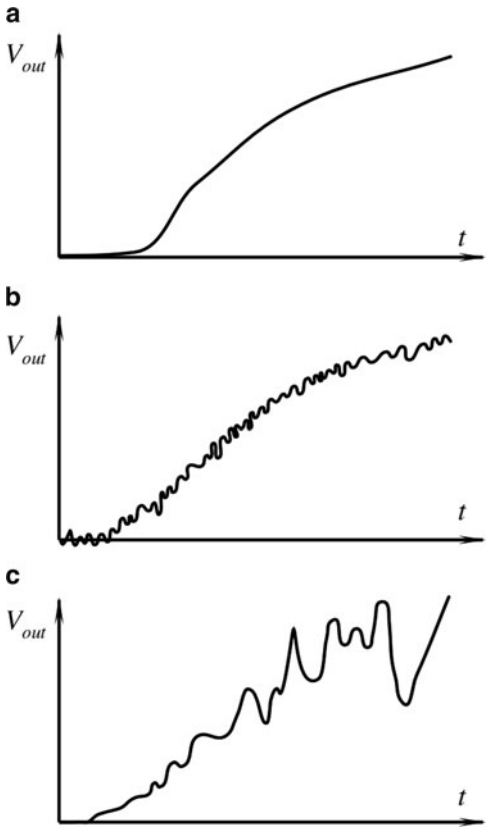

Fig. 5.45 Types of noise noise-free signal (**a**); additive noise (**b**); multiplicative noise (**c**)

where $N(t)$ is a function of noise. An example of such noise is shown in Fig. 5.45c. Multiplicative noise at the output disappears or becomes small (it also becomes additive) when the signal magnitude nears zero. Multiplicative noise grows together with the signal's V_s magnitude. As its name implies, multiplicative noise is a result of multiplication (which is essentially a nonlinear operation) of two values where one is a useful signal and the other is a noise dependent spurious signal.

To improve noise stability against transmitted additive noise, quite often sensors are combined in pairs, that is, they are fabricated in a dual form whose output signals are subtracted from one another (Fig. 5.46). This method is called a differential technique (see also Sect. 5.10). One sensor of the pair (it is called the main sensor) is subjected to a stimulus of interest s_1, while the other (reference) is shielded from stimulus perception. Since additive noise is specific for the linear or quasilinear sensors and circuits, the reference sensor does not have to be subjected to any particular stimulus. Often, it may be equal to zero. It is anticipated that both sensors are subjected to identical transmitted noise (nose generated inside the sensor cannot be cancelled by a differential technique), which it is said, is

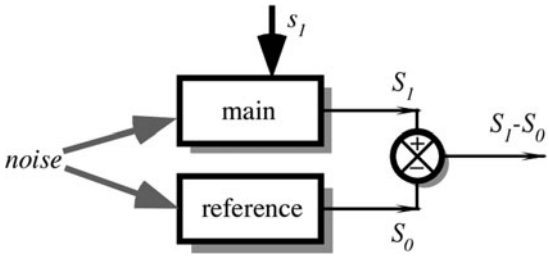

Fig. 5.46 Differential technique

a common-mode noise. This means that noisy effects at both sensors are in-phase and have the same magnitude. If both sensors are identically influenced by common-mode spurious stimuli, the subtraction removes the noise component. Such a sensor is often called either a dual or a differential sensor. The quality of noise rejection is described by a number which is called the common-mode rejection ratio (CMRR):

$$\text{CMMR} = 0.5\frac{S_1 + S_0}{S_1 - S_0}, \tag{5.56}$$

where S_1 and S_0 are output signals from the main and reference sensors, respectively. CMRR may depend on magnitude of stimuli and usually becomes smaller at greater input signals. The ratio shows how many times stronger the actual stimulus will be represented at the output, with respect to a common mode noise having the same magnitude. The value of the CMRR is a measure of the sensor's symmetry. To be an effective means of noise reduction, both sensors must be positioned as close as possible to each other, they must be very identical and subjected to the same environmental conditions. Also, it is very important that the reference sensor is reliably shielded from the actual stimulus, otherwise the combined differential response will be diminished.

To reduce transmitted multiplicative noise, a ratiometric technique is quite powerful (see Sect. 5.9 for circuit description). Its principle is quite simple. The sensor is fabricated in a dual form where one part is subjected to the stimulus of interest and both parts are subjected to the same environmental conditions, which may cause transmitted multiplicative noise. The second sensor is called reference because a constant environmentally stable reference stimulus s_o is applied to its input. For example, the output voltage of a sensor in a narrow temperature range may be approximated by equation

$$V_1 \approx [1 + \alpha(T - T_0)]f(s_1), \tag{5.57}$$

where α is the temperature coefficient of the sensor's transfer function, T is the temperature, and T_0 is the temperature at calibration. The reference sensor whose reference input is s_0 generates voltage

$$V_0 \approx [1 + \alpha(T - T_0)]f(s_0). \tag{5.58}$$

We consider ambient temperature as a transmitted multiplicative noise which affects both sensors in the same way. Taking ratio of the above equations we arrive at

$$\frac{V_1}{V_0} = \frac{1}{f(s_0)} f(s_1). \tag{5.59}$$

Since $f(s_0)$ is constant, the ratio is not temperature-dependent. It should be emphasized however that the ratiometric technique is useful only when the anticipated noise has a multiplicative nature, while a differential technique works only for additive transmitted noise. Neither technique is useful for inherent noise, which is generated internally in sensors and circuits.

While inherent noise is mostly Gaussian, the transmitted noise is usually less suitable for conventional statistical description. Transmitted noise may be periodic, irregularly recurring, or essentially random, and it ordinarily may be reduced substantially by taking precautions to minimize electrostatic and electromagnetic pickup from power sources at line frequencies and their harmonics, radio broadcast stations, arcing of mechanical switches, and current and voltage spikes resulting from switching in reactive (having inductance and capacitance) circuits. Such precautions may include filtering, decoupling, shielding of leads and components, use of guarding potentials, elimination of ground loops, physical reorientation of leads, components and wires, use of damping diodes across relay coils and electric motors, choice of low impedances where possible, and choice of power supply and references having low noise. Transmitted noise from vibration may be reduced by proper mechanical design. A list outlining some of the sources of transmitted noise, their typical magnitudes, and some ways of dealing with them is shown in Table 5.4.

Table 5.4 Typical sources of transmitted noise (adapted from [13])

External source	Typical magnitude	Typical cure
60/50 Hz power	100 pA	Shielding; attention to ground loops; isolated power supply
120/100 Hz supply ripple	3 μV	Supply filtering
180/150 Hz magnetic pickup from saturated 60/50 Hz transformers	0.5 μV	Reorientation of components
Radio broadcast stations	1 mV	Shielding
Switch-arcing	1 mV	Filtering of 5 to 100 MHz components; attention to ground loops and shielding
Vibration	10 pA (10–100 Hz)	Proper attention to mechanical coupling; elimination of leads with large voltages near input terminals and sensors
Cable vibration	100 pA	Use a low noise (carbon coated dielectric) cable
Circuit boards	$0.01 - 10\,\text{pA}/\sqrt{\text{Hz}}$ below 10 Hz	Clean board thoroughly; use Teflon insulation where needed and guard well

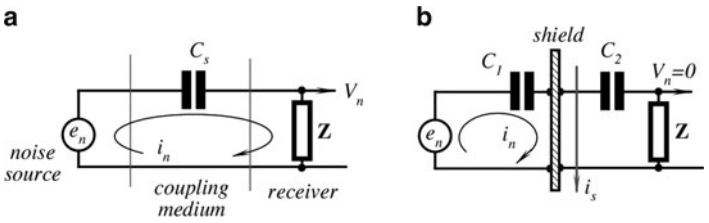

Fig. 5.47 Capacitive coupling (**a**) and electric shield (**b**)

The most frequent channel for the coupling of electrical noise is a "parasitic" capacitance. Such a coupling exists everywhere. Any object is capacitively coupled to another object. For instance, a human standing on isolated earth develops a capacitance to ground on the order of 700 pF, electrical connectors have a pin-to-pin capacitance of about 2 pF, an optoisolator has an emitter-detector capacitance of about 2 pF. Figure 5.47a shows that an electrical noise source is connected to the sensor's internal impedance **Z** through a coupling capacitance C_s. That impedance may be a simple resistance or a combination of resistors, capacitors, inductors, and nonlinear elements, like diodes. Voltage across the impedance is a direct result of the change rate in the noise signal, the value of coupling capacitance C_s and impedance **Z**. For instance, a pyroelectric detector may have an internal impedance, which is equivalent to a parallel connection of a 30 pf capacitor and a 50 GΩ resistor. The sensor may be coupled through just 1 pf to a moving person who has the surface electrostatic charge on the body resulting in static voltage of 1,000 V. If we assume that the main frequency of human movement is 1 Hz, the sensor would pickup the electrostatic interference of about 30 V. This is 3 to 5 orders of magnitude higher than the sensor would normally produce in response to thermal radiation received from the human body.

Since some sensors and virtually all electronic circuits have nonlinearities, high-frequency interference signals, generally called RFI (radiofrequency interference) or EMI (electromagnetic interferences), may be rectified and appear at the output as a dc or slow changing voltage.

5.13.3 Electric Shielding

Interferences attributed to electric fields can be significantly reduced by appropriate shielding of the sensor and circuit, especially of high impedance and nonlinear components. Each shielding problem must be analyzed separately and carefully. It is very important to identify the noise source and how it is coupled to the circuit. Improper shielding and guarding may only make matters worse or create a new problem.

A shielding serves two purposes [14]. First, it confines noise to a small region. This will prevent noise from getting into nearby circuits. However, the problem

with such shields is that the noise captured by the shield can still cause problems if the return path that the noise takes is not carefully planned and implemented by an understanding of the ground system and making the connections correctly.

Second, if noise is present in the circuit, shields can be placed around critical parts to prevent the noise from getting into sensitive portions of the detectors and circuits. These shields may consist of metal boxes around circuit regions or cables with shields around the center conductors.

As it was shown in Sect. 3.1, the noise that resulted from the electric fields can be well controlled by metal enclosures because charge q cannot exist on the interior of a closed conductive surface. Coupling by a mutual, or stray, capacitance can be modeled by a circuit shown in Fig. 5.47a. Here e_n is a noise source. It may be some kind of a part or component whose electric potential varies. C_s is the stray capacitance (having impedance Z_s at a particular frequency) between the noise source and the circuit impedance Z, which acts as a receiver of the noise. Voltage V_n is a result of the capacitive coupling. A noise current is defined as

$$i_n = \frac{V_n}{Z + Z_s}, \tag{5.60}$$

and actually produces noise voltage

$$V_n = \frac{e_n}{\left(1 + \frac{Z_c}{Z}\right)}. \tag{5.61}$$

For example, if $C_s = 2.5$pf, $Z = 10$ kΩ (resistor) and $e_n = 100$ mV, at 1.3 MHz, the output noise will be 20 mV.

One might think that 1.3 MHz noise is relatively easy to filter out from low-frequency signals produced by a sensor. In reality, it cannot be done, because many sensors and, especially the front stages of the amplifiers, contain nonlinear components (*pn*-semiconductor junctions), which act as rectifiers. As a result, the spectrum of high-frequency noise shifts into a low-frequency region, making the noise signal similar to voltage produced by a sensor.

When a shield is added, the change to the situation is shown in Fig. 5.47b. With the assumption that the shield has zero impedance, the noise current at the left side will be $i_n = e_n/Z_c$. On the other side of the shield, noise current will be essentially zero since there is no driving source at the right side of the circuit. Subsequently, the noise voltage over the receiving impedance will also be zero and the sensitive circuit becomes effectively shielded from the noise source. One must be careful, however, that there is no significant currents i_s flow over the shield. Coupled with the shield resistance, these may generate additional noise. There are several practical rules that must be observed when applying electrostatic shields.

- An electrostatic shield, to be effective, should be connected to the reference potential of any circuitry contained within the shield. If the signal is connected to a ground (chassis of the frame or to earth), the shield must be connected to that ground. Grounding of shield is useless if the signal is not returned to the ground.

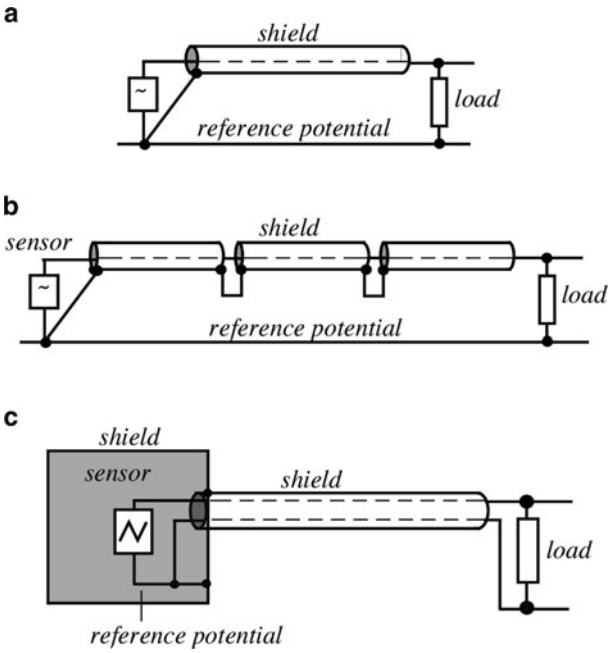

Fig. 5.48 Connections of an input cable to a reference potential

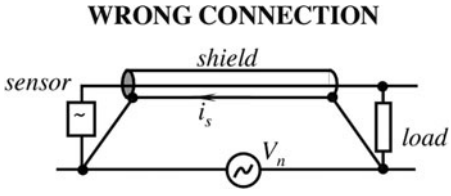

Fig. 5.49 Cable shield is erroneously grounded at both ends

- If a shielding cable is used, its shield must be connected to the signal referenced node at the signal source side (Fig. 5.48a).
- If the shield is split into sections, as might occur if connectors are used, the shield for each segment must be tied to those for the adjoining segments, and ultimately connected only to the signal referenced node (Fig. 5.49b).
- The number of separate shields required in a data acquisition system is equal to the number of independent signals that are being measured. Each signal should have its own shield, with no connection to other shields in the system, unless they share a common reference potential (signal "ground"). In that case all connections must be made by a separate jumping wire connected to each shield at a single point.

- A shield must be grounded only at one point, preferably next to the sensor. A shielded cable must *never* be grounded at both ends (Fig. 5.49). The potential difference (V_n) between two "grounds" will cause shield current i_s to flow, which may induce a noise voltage into the center conductor via magnetic coupling.
- If a sensor is enclosed into a shield box and data are transmitted via a shielded cable (Fig. 5.48c), the cable shield must be connected to the box. It is a good practice to use a separate conductor for the reference potential ("ground") inside the shield, and not to use the shield for any other purposes except shielding: do not allow shield current to exist.
- Never allow the shield to be at any potential with respect to the reference potential (except in case of driven shields as shown in Fig. 5.4b). The shield voltage couples to the center conductor (or conductors) via a cable capacitance.
- Connect shields to a ground via short wires to minimize inductance. This is especially important when both analog and digital signals are transmitted.

5.13.4 Bypass Capacitors

The bypass capacitors are used to maintain low power supply impedance at the point of a load. Parasitic resistance and inductance in supply lines mean that the power supply impedance can be quite high. As the frequency goes up, the inductive parasitic becomes troublesome and may result in the circuit oscillation or ringing effects. Even if the circuit operates at lower frequencies, the bypass capacitors are still important as high-frequency noise may be transmitted to the circuit and power supply conductors from external sources, for instance radio stations. At high frequencies, no power supply or regulator has zero output impedance. What type of capacitor to use is determined by the application, frequency range of the circuit, cost, board space, and some other considerations. To select a bypass capacitor, one must remember that a practical capacitor at high frequencies may be far away from the idealized capacitor, which is described in textbooks.

A generalized equivalent circuit of a capacitor is shown in Fig. 5.50. It is comprised of a nominal capacitance C, leakage resistance r_l, lead inductances L, and resistances R. Further, it includes dielectric absorption terms r and c_a, which are manifested in capacitor's "memory." In many interface circuits, especially amplifiers, analog integrators and current (charge)-to-voltage converters, dielectric

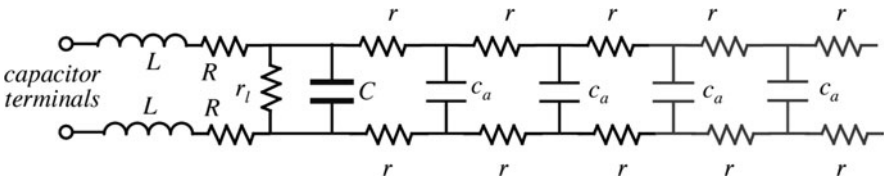

Fig. 5.50 Equivalent circuit of a capacitor

absorption is a major cause for errors. In such circuits, film capacitors should be used whenever possible.

In bypass applications, r_l and dielectric absorption are second order terms but series R and L are of importance. They limit the capacitor's ability to damp transients and maintain a low power supply output impedance. Often, bypass capacitors must be of large values (10 μF or more) so they can absorb longer transients, thus electrolytic capacitors are often employed. Unfortunately, these capacitors have large series R and L. Usually, tantalum capacitors offer better results, however, a combination of aluminum electrolytic with nonpolarized (ceramic or film) capacitors may offer even further improvement. Nowadays, high-volume ceramic capacitors are available for low price. A combination of wrong types of bypass capacitors may lead to ringing, oscillation and crosstalk between data communication channels. The best way to specify a correct combination of bypass capacitors is to first try them on a breadboard.

5.13.5 Magnetic Shielding

Proper shielding may dramatically reduce noise resulting from electrostatic and electrical fields. Unfortunately, it is much more difficult to shield against magnetic fields because it penetrates conducting materials. A typical shield placed around a conductor and grounded at one end has little if any effect on the magnetically induced voltage in that conductor. As magnetic field B_o penetrates the shield, its amplitude drops exponentially (Fig. 5.51b). The skin depth δ of the shield is the depth required for the field attenuation by 37% of that in the air. Table 5.5 lists typical values of δ for several materials at different frequencies. At high frequencies, any material from the list may be used for effective shielding, however at a lower range steel yields a much better performance.

For improving low-frequency magnetic field shielding, a shield consisting of a high-permeability magnetic material (e.g., mumetal) should be considered. However, the mumetal effectiveness drops at higher frequencies and strong magnetic fields. An effective magnetic shielding can be accomplished with thick steel shields

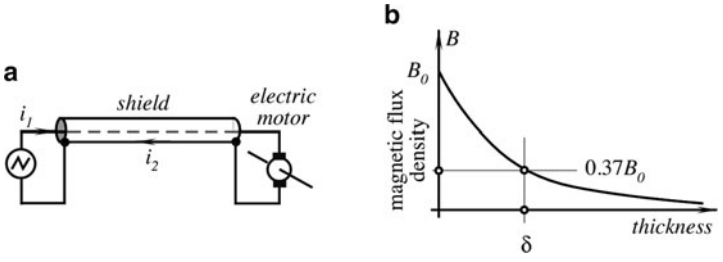

Fig. 5.51 Reduction of a transmitted magnetic noise by powering a load device through a coaxial cable (**a**); Magnetic shielding improves with the thickness of the shield (**b**)

Table 5.5 Skin depth, δ, in mm versus frequency (adapted from [15])

Frequency	Copper	Aluminum	Steel
60 Hz	8.5	10.9	0.86
100 Hz	6.6	8.5	0.66
1 kHz	2.1	2.7	0.20
10 kHz	0.66	0.84	0.08
100 kHz	0.2	0.3	0.02
1 MHz	0.08	0.08	0.008

at higher frequencies. Since magnetic shielding is very difficult, the most effective approach at low frequencies is to minimize the strength of magnetic fields, minimize the magnetic loop area at the receiving end, and selecting the optimal geometry of conductors. Some useful practical guidelines are as follows:

- Locate the receiving circuit as far as possible from the source of the magnetic field.
- Avoid running wires parallel to the magnetic field; instead, cross the magnetic field at right angles.
- Shield the magnetic field with an appropriate material for the frequency and strength.
- Use a twisted pair of wires for conductors carrying the high-level current that is the source of the magnetic field. If the currents in the two wires are equal and opposite, the net field in any direction over each cycle of twist will be zero. For this arrangement to work, none of the current can be shared with another conductor, for example, a ground plane, which may result in ground loops.
- Use a shielded cable with the high-level source circuit's return current carried by the shield (Fig. 5.51a). If the shield current i_2 is equal and opposite to that of the center conductor i_1, the center conductor field and the shield field will cancel, producing a zero net field. This case seems a violation of a rule "no shield currents" for the receiver's circuit, however, the shielded cable here is not used to electrostatically shield the center conductor. Instead, the geometry produces a cancellation of the magnetic field which is generated by a current supplied to a "current-hungry" device (an electric motor in this example).
- Since magnetically induced noise depends on the area of the receiver loop, the induced voltage due to magnetic coupling can be reduced by making the loop's area smaller.

What is the receiver's loop? Figure 5.52 shows a sensor, which is connected to the load circuit via two conductors having length L and separated by distance D. The rectangular circuit forms a loop area $a = L{\cdot}D$. The voltage induced in series with the loop is proportional to the area and cosine of its angle to the field. Thus, to minimize noise, the loop should be oriented at right angles to the field, and its area should be minimized.

The area can be decreased by reducing the length of the conductors and/or decreasing the distance between the conductors. This is easily accomplished with a twisted pair, or at least with a tightly cabled pair of conductors. It is a good practice to pair the conductors so that the circuit wire and its return path will always

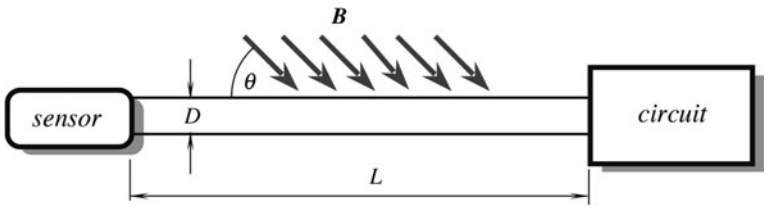

Fig. 5.52 Receiver's loop is formed by long conductors

be together. This requirement shall not be overlooked. For instance, if wires are correctly positioned by a designer, a service technician may reposition them during the repair work. A new wire location may create a disastrous noise level. Hence, a general rule is to know the area and orientation of the wires and permanently secure the wiring.

5.13.6 Mechanical Noise

Vibration and acceleration effects are also sources of transmitted noise in sensors, which otherwise should beimmune to them. These effects may alter transfer characteristics (multiplicative noise) or the sensor may generate spurious signals (additive noise). If a sensor incorporates certain mechanical elements, vibration along some axes with a given frequency and amplitude may cause resonant effects. For some sensors, an acceleration is a source of noise. For instance, pyroelectric detectors possess piezoelectric properties. The main function of a pyroelectric detector is to respond to thermal gradients. However, such environmental mechanical factors as a fast changing air pressure, strong wind or structural vibrations cause the sensor to respond with output signals, which often are indistinguishable from responses to normal stimuli. If this is the case, a differential noise cancellation may be quite efficient (see Sect. 5.8).

5.13.7 Ground Planes

For many years, ground planes have been known to electronic engineers and printed circuit designers as a "mystical and ill-defined" cure for spurious circuit operation [16]. Ground planes are primarily useful for minimizing circuit inductance. They do this by utilizing the basic magnetic theory. Current flowing in a wire produces an associated magnetic field (Sect. 3.3.1). The field's strength is proportional to the current i and inversely related to the distance r from the conductor:

$$B = \frac{\mu_0 i}{2\pi r}. \tag{5.62}$$

Thus, we can imagine a current-carrying wire surrounded by a magnetic field. Wire inductance is defined as energy stored in the field setup by the wire's current. To compute the wire's inductance requires integrating the field over the wire's length and the total area of the field. This implies integrating on the radius from the wire surface to infinity. However, if two wires carrying the same current in opposite directions are in close proximity, their magnetic fields are canceled. In this case, the virtual wire inductance is much smaller. An opposite flowing current is called *return current*. This is the underlying reason for ground planes. A ground plane provides a return path directly under the signal carrying conductor through which return current can flow. Return current has a direct path to ground, regardless of the number of branches associated with the conductor. Currents will always flow through the return path of the lowest impedance. In a properly designed ground plane, this path is directly under the signal conductor.

In practical circuits, a ground plane is one side of the board and the signal conductors are on the other. In the multilayer boards, a ground plane is usually sandwiched between two or more conductor planes. Aside from minimizing parasitic inductance, ground planes have additional benefits. Their flat surface minimizes resistive losses due to "skin effect" (ac current travel along a conductor's surface). Additionally, they aid the circuit's high-frequency stability by referring stray capacitance to the ground. Even though ground planes are very beneficial for digital circuits using them for current return of analog sensor signals are dangerous likely digital currents in a ground will create strong interferences in the analog part of the circuit.

Some practical suggestions:

- Make ground planes of as much area as possible on the components side (or inside for the multilayer boards). Maximize the area especially under traces that operate with high frequency or digital signals.
- Mount components that conduct fast transient currents (terminal resistors, ICs, transistors, decoupling capacitors, etc.) as close to the board as possible.
- Wherever a common ground reference potential is required, use separate conductors for the reference potential and connect them all to the ground plane at a common point to avoid voltage drops due to ground currents.
- Use separate nonoverlapping ground planes for digital and analog sections of the circuit board and connect them at one point only at the power supply terminals.
- Keep the trace length short. Inductance varies directly with length and no ground plane will achieve perfect cancellation.

5.13.8 Ground Loops and Ground Isolation

When a circuit is used for low-level input signals, a circuit itself may generate enough noise and interferences to present a substantial problem for accuracy. Sometimes, when a circuit is correctly designed on paper, a bench breadboard shows quite a satisfactory performance, however, when a production prototype with

the printed circuit board is tested, the accuracy requirement is not met. A difference between a bread board and PC board prototypes may be in the physical layout of conductors. Usually, conductors between electronic components are quite specific; they may connect a capacitor to a resistor, a gate of a JFET transistor to the output of an operational amplifier, etc. However, there are at least two conductors which, in most cases, are common for the majority of the electronic circuit. These are the power supply bus and the ground bus. Both of them may carry undesirable signals from one part of the circuit to another, specifically, they may couple strong output signals to the sensors and input stages.

A power supply bus carries supply currents to all stages. A ground bus also carries supply currents, but, in addition, it is often used to establish a reference base for an electrical signal. For any measurement circuit cleanliness of a reference base is essential. Interaction of the two functions (power supply and reference) may lead to a problem which is known as *ground loop*. We illustrate it in Fig. 5.53a where a sensor is connected to a positive input of an amplifier which may have a substantial gain. The amplifier is connected to the power supply and draws current i which is returned to the ground bus as i'. A sensor generates voltage V_s, which is fed to the positive input of the amplifier. A ground wire is connected to the circuit at point a, right next to the sensor's terminal. A circuit has no visible error sources, nevertheless, the output voltage contain substantial error. A noise source is developed in a wrong connection of ground wires. Figure 5.53b shows that the ground conductor is not ideal. It may have some finite resistance R_g and inductance L_g. In this example, supply current while returning to the battery from the amplifier passes through the ground bus between points b and a resulting in voltage drop V_g. This drop, however small, may be comparable with the signal produced by the sensor. It should be noted that voltage V_g is serially connected with the sensor and is directly applied to the amplifier's input. In other words, the sensor is not referenced to a clean ground. Ground currents may also contain high-frequency components, then the bus inductance will produce quite strong spurious high-frequency signals, which not only add

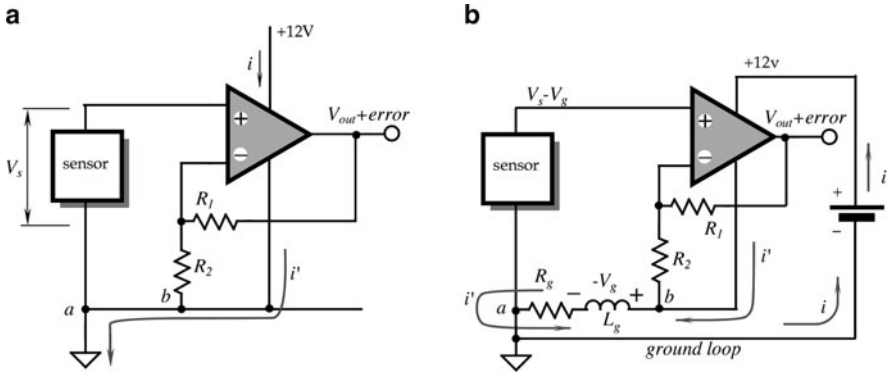

Fig. 5.53 Wrong connection of a ground terminal to a circuit (**a**); Path of a supply current through the ground conductors (**b**)

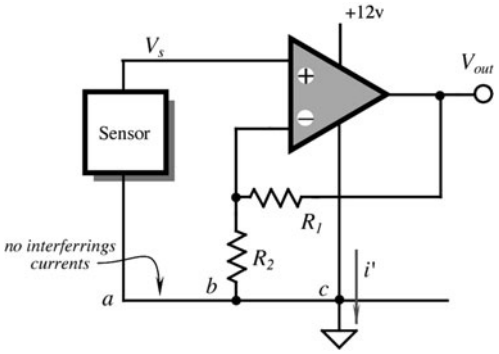

Fig. 5.54 Correct grounding of a sensor and interface circuit

noise to the sensor, but may cause circuit instability as well. For example, let us consider a thermopile sensor, which produces voltage corresponding to 100 μV/°C of the object's temperature. A low-noise amplifier has quiescent current, $i = 2$ mA, which passes through the ground loop having resistance $R_g = 0.2\ \Omega$. The ground loop voltage $V_g = iR_g = 0.4$ mV corresponds to an error of −4°C!

The cure is usually quite simple: ground loops must be broken. The most critical rule of the circuit board design: never use the same conductor for the reference potential and power supply currents. A circuit designer should always separate a reference ground from the current carrying grounds, especially serving digital devices. Thus, it is advisable to have at least three grounds: reference, analog, and digital.

The reference ground shall be used only for connecting the sensor components that produce low level input signals, all analog amplifier input components that need be referenced to a ground potential, and the reference input of an A/D converter. The analog ground shall be used exclusively for return currents from the analog interface circuits. And the digital ground shall be used only for binary signals, like microprocessors, digital gates, etc. There may be a need for additional "grounds," for example those that carry relatively strong currents, especially containing high-frequency signals (LEDs, relays, motors, heaters, etc.). Figure 5.54 shows that moving the ground connection from the sensor's point *a* to the power terminal point *c* prevents formation of spurious voltages across the ground conductor connected to the sensor and a feedback resistor R_2.

A rule of thumb is to join all "grounds" on a circuit board only at one point, preferably at the power source. Grounding at two or more spots may form ground loops, which often is very difficult to diagnose.

5.13.9 Seebeck Noise

This noise is a result of the Seebeck effect (Sect. 3.9), which is manifested as the generation of an electromotive force (e.m.f.) when two dissimilar metals are joined

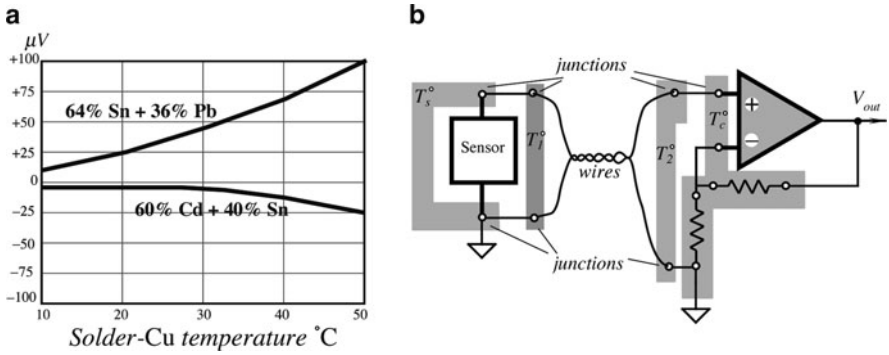

Fig. 5.55 Seebeck *e.m.f.* developed by solder-copper joints (**a**) (adapted from [17]); Maintaining joints at the same temperature reduces Seebeck noise (**b**)

together. The Seebeck e.m.f. is small and for many sensors may be simply ignored. However, when absolute accuracy on the order of 10–100 μV is required, that noise must be taken into account. The connection of two dissimilar metals produces a temperature sensor. However, when temperature sensing is not a desired function, a thermally induced *e.m.f.* is a spurious signal. In electronic circuits, connection of dissimilar metals can be found everywhere: connectors, switches, relay contacts, sockets, wires, etc. For instance, the copper PC board cladding connected to kovar™[5] input pins of an integrated circuit creates an offset voltage of 40 μV·Δ*T* where Δ*T* is the temperature gradient in °C between two dissimilar metal contacts on the board. The common lead-tin solder, when used with the copper cladding creates a thermoelectric voltage between 1 and 3 μV/°C. There are special cadmium-tin solders available to reduce these spurious signals down to 0.3 μV/°C. Figure 5.55a shows Seebeck *e.m.f.* for two types of solder. Connection of two identical wires fabricated by different manufacturers may result in voltage having slope on the order of 200 nV/°C.

In many cases, Seebeck e.m.f. may be eliminated by a proper circuit layout and thermal balancing. It is a good practice to limit the number of junctions between the sensor and the front stage of the interface circuit. Avoid connectors, sockets, switches and other potential sources of *e.m.f.* to the extent possible. In some cases this will not be possible. In these instances, attempt to balance the number and type of junctions in the circuit's front stage so that differential cancellations occur. Doing this may involve deliberately creating and introducing junctions to offset necessary junctions. Junctions, which intent to produce cancellations, must be maintained at the same temperature. Figure 5.55b shows a remote sensor connection to an amplifier where the sensor junctions, input terminal junctions, and amplifier components junctions are all maintained while at different but properly arranged temperatures. Such thermally balanced junctions must be maintained at a

[5]Trademark of Westinghouse Electric Corp.

close physical proximity and preferably on common heat sinks. Air drafts and temperature gradients in the circuit boards and sensor enclosures must be avoided.

5.14 Calibration

Many sensors have natural manufacturer's tolerances that move their transfer functions beyond the acceptable accuracy limits. We discussed the calibrating principles in detail in Chap. 2. Now, let us briefly examine some electrical techniques that can be employed in the interface circuits to trim the overall transfer functions including both the sensor and the interface circuit. The non-exhaustive list of techniques is as follows:

1. A "trimpot" or "trimcap," a variable resistor or capacitor to adjust the sensor's internal resistance or capacitance. This is the oldest and most traditional method. Figure 5.56a illustrates use of the trimpot R_p in the resistive Wheatstone bridge. Presently, the trimming resistors may have digital rather than mechanical adjustments. In digital potentiometers (for example, from www.maxim-ic.com) an 8-bit code can set the resistance value and be changed at any time. A more permanent nonmechanical trimming of a resistor can be done by using methods developed by Microbridge Technologies, Inc. (www.mbridgetech.com). The trimming resistor (given by the manufacturer a strange name "rejustor") is a semiconductor whose ohmic resistance can be modified by heating pulses applied through separate terminals (Fig. 5.56c) from a special trimming equipment. Resistance can be reduced from the nominal as-manufactured value by at least 30% and then can be adjusted up or down within the active range. The programmed resistance value is stored in the physical properties of the semiconductor and retained until it is reprogrammed. No power is required to maintain the resistance value; it is strictly a passive circuit element.

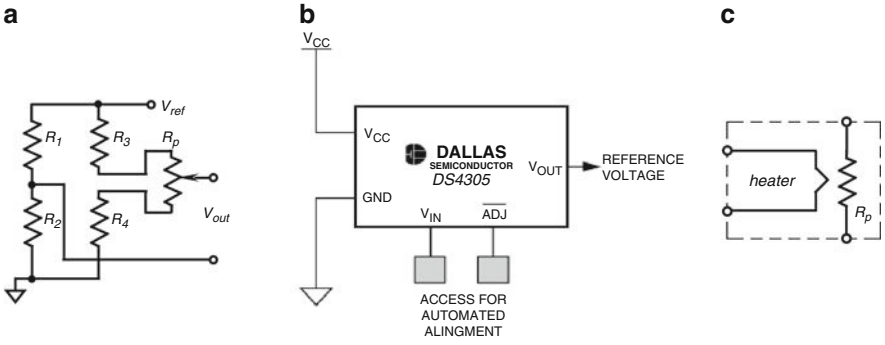

Fig. 5.56 Trimming circuits resistive bridge circuit with a trimming resistor (**a**); digitally adjustable voltage reference (**b**); and "rejustor," thermally trimmed resistor (**c**)

2. A matching, that is, selectable resistor or capacitor to match the corresponding resistance or capacitance of the sensor.
3. Adjust the sensor's reference signal. This can be done by modifying one of the excitation parameters of the active sensor's excitation signal, for example, amplitude or frequency. Alternatively, a reference voltage can be adjusted at the A/D converter. Figure 5.56b illustrates a digitally programmable voltage reference.
4. Digital code that is stored in a nonvolatile memory of the microprocessor and during every measurement is used to correct the sensor's response. The code is created during the calibration procedure. It may include numerous sensors characteristics, for example, coefficients for a polynomial approximation.

5.15 Batteries for Low-Power Sensors

Modern development of integrated sensors and need for long-term remote monitoring and data acquisition demand use of reliable and high-energy density power sources. History of battery development goes back to Volta and shows a remarkable progress during last decades. Well-known old electrochemical power sources improve dramatically. Examples are C-Zn, alkaline, Zn-air, NiCd, and lead-acid batteries. Nowadays, newer systems such as secondary Zn-air, Ni-metal-hydride, and especially lithium batteries are growing in use as new devices are designed around their higher voltage and superior shelf life. The Li-MnO_2 system dominates the commercial market where they range from miniature flat cells to "D" size cells.

All batteries can be divided into two groups: primary – single use devices and secondary (rechargeable) – multiple use devices.

Often, batteries are characterized by energy per unit weight, however, for miniature sensor applications energy per unit volume often becomes more critical (see in Appendix Table A-20).

In general, the energy delivered by a battery depends upon the rate at which power is withdrawn. Typically, as the current is increased the amount of energy delivered is decreased. Battery energy and power are also affected by construction of battery, the size, and the duty cycle of current delivery. Manufacturers usually specify batteries as ampere-hours or watt-hours when discharged at a specific rate to a specific voltage cut-off.

If the battery capacitance is C (in mA·hour) and the average current drain is I (mA), the time of a battery discharge (lifetime for a primary cell) is defined as

$$t = \frac{C}{In} \tag{5.63}$$

where n is a duty cycle. For instance, if the battery is rated as having capacity of 100 mA h, the circuit operating current consumption is about 5 mA and the circuit works only 3 min every hour (duty cycle is 3/60), the battery will last approximately for

$$t = \frac{C}{In} = \frac{100}{5\frac{3}{60}} = 400\,\text{h}$$

Yet, the manufacturer's specification shall be used with a grain of salt and only as a guideline, because the specified discharge rate is rarely coincides with the actual power consumption. It is highly recommended to determine battery life experimentally, rather than rely on the calculation. When designing the electronic circuit, its power consumption shall be determined during various operating modes and over the operating temperature range. Then, these values of power consumption shall be used in simulation of the battery load to determine the useful life with a circuit-specific cut-off voltage in mind. Sometimes, a circuit draws high currents during short times (pulse mode) and the battery ability to deliver such pulse current should be evaluated. If a battery cannot deliver high pulse current, a parallel electrolytic capacitor serving as a storage tank may be considered.

It should be noted that the accelerated life tests of a battery shall be used with caution, since as it was noted above, the useful capacity of a battery greatly depends on the load, operational current profile, and a duty cycle.

5.15.1 Primary Cells

The construction of a battery cell determines its performance and cost. Most primary cells (disposable batteries) employ single thick electrodes arranges in parallel or concentric configuration and aqueous electrolytes. Most small secondary cells (rechargeable batteries) are designed differently; they use "wound" or "jelly roll" construction, in which long thin electrodes are wound into a cylinder and placed into a metal container. This results in a higher power density, but with decreased energy density and higher cost. Due to low conductivity of electrolytes, many lithium primary cells also use "wound" construction [18].

5.15.1.1 Leclanche (Carbon-Zinc) Batteries

These batteries use zinc as anode. There are two types of them. One uses natural manganese dioxide as the cathode with ammonium chloride electrolyte. A "premium" version uses electrolytic manganese dioxide as the cathode and a zinc chloride electrolyte. These batteries are still the most popular world-wide, especially in the Orient, being produced by over 200 manufacturers. Their use is about equal to that of the alkaline in Europe but is only near 25% of alkaline in the U.S.A. These batteries are preferred when high power density is not required, shelf life is not critical, but the low cost is a dominating factor.

5.15.1.2 Alkaline Manganese Batteries

Demand for these batteries grew significantly, especially after a major improvement—elimination of mercury from the zinc anode. The alkaline batteries are capable of delivering high currents, have improved power/density ratio and at least 5 years of shelf life (Table A.20)

5.15.1.3 Primary Lithium Batteries

Most of these batteries are being produced in Japan and China. The popularity of lithium-manganese dioxide cells grows rapidly thanks to their higher operating voltage, wide range of sizes and capacities, and excellent shelf life (Table A.21). Lithium iodine cells have very high energy density and allow up to 10 years of operation in a pacemaker (implantable heart rate controller). However, these batteries are designed with a low conductivity solid-state electrolyte and allow operation with very low current drain (in the order of microamperes), which in cases when passive sensors are employed often is quite sufficient.

Amount of lithium is batteries is quite small, because just 1 g is sufficient for producing capacity of 3.86Ah. Lithium cells are exempt from environmental regulations, but still are considered hazardous because of their flammability.

5.15.2 Secondary Cells

Secondary cells (Tables A.22 and A.23) are rechargeable batteries.

Sealed lead acid batteries offer small size at large capacities and allow about 200 cycles of life at discharge times as short as 1 h. The main advantages of these cells are low initial cost, low self-discharge, an ability to support heavy loads, to withstand harsh environments. Besides, these batteries have long life. The disadvantages include relatively large size and weight as well as potential environmental hazard due to presence of lead and sulfuric acid.

Sealed nickel-cadmium (NiCd) and nickel-metal hydrate (Ni-MH) are the most widely used secondary cells, being produced at volumes over 1 billion cells per year. Typical capacity for a "AA" cell is about 800 mAh and even higher from some manufacturers. This is possible thanks to use a high-porosity nickel foam or felt instead of traditional sintered nickel as carrier for the active materials. The NiCd cells are quite tolerant of overcharge and overdischarge. An interesting property of NiCd is that charging is endothermic process, which is the battery absorbs heat, while other batteries warm up when charging. Cadmium, however, presents potential environmental problem. Bi-MH and modern NiCd do not exhibit "memory" effect, that is, partial discharge does not influence their ability to fully recharge. The nickel-metal hydrate cells is nearly direct replacement for NiCd, yet they yield better capacity, but have somewhat poorer self-discharge.

A lithium polymer battery contain a nonliquid electrolyte, which makes it a solid-state battery. This allows to fabricate it in any size and shape, however, these batteries are most expensive.

Rechargeable alkaline batteries have low cost and good power density. However, their life cycles are quite low.

References

1. Widlar RJ (1980) Working with high impedance Op Amps, AN24, *Linear Application Handbook*. National Semiconductor
2. Pease RA (1983) Improve circuit performance with a 1-op-amp current pump. *EDN*, 85–90, Jan. 20
3. Sheingold DH (ed) (1986) *Analog-Digital Conversion Handbook*. 3rd ed., Prentice-Hall, Englewood Cliffs, NJ
4. Williams J (1990) Some techniques for direct digitization of transducer outputs, AN7, *Linear Technology Application Handbook*
5. Park YE, Wise KD (1983) An MOS switched-capacitor readout amplifier for capacitive pressure sensors. *IEEE Custom IC Conf* 380–384
6. Stafford KR, Gray PR, Blanchard RA (1974) A complete monolithic sample/hold amplifier. *IEEE J Solid-State Circuits* 9:381–387
7. Cho ST, Wise KD (1991) A self-testing ultrasensitive silicon microflow sensor. *Sensor Expo Proceedings*, 208B-1
8. Weatherwax S (1991) Understanding constant voltage and constant current excitation for pressure sensors. *SenSym Solid-State Sensor Handbook*. ©Sensym, Inc.
9. Coats MR (1991) New technology two-wire transmitters. *Sensors* 8(1)
10. Sheingold DH (ed) (1974) *Nonlinear Circuits Handbook*. Analog Devices, Inc. Northwood, MA
11. Johnson JB (1928) Thermal agitation of electricity in conductors. *Phys Rev*
12. The Best of Analog Dialogue. © Analog Devices, Inc. 1991
13. Rich A (1991) Shielding and guarding. *Best of Analog Dialogue*, ©Analog Devices, Inc.
14. Ott HW (1976) *Noise Reduction Techniques in Electronic Systems*. Wiley, New York
15. Williams J (1990) High speed comparator techniques, AN13. *Linear Applications Handbook*. ©Linear Technology Corp.
16. Pascoe G (1977) The choice of solders for high-gain devices. *New Electronics* (U.K.), Feb. 6
17. Powers RA (1995) Batteries for low power electronics. *Proc IEEE* 83(4):687–693
18. AVR121 (2005) Enhancing ADC resolution by oversampling. Atmel Application Note 8003A-AVR-09/05
19. Bell DA (1981) *Solid State Pulse Circuits*. 2nd ed. Reston Publishing Company, Inc., Reston, VA

Chapter 6
Occupancy and Motion Detectors

"Never confuse motion with action".

- Ernest Hemingway

September 11 has changed the way people think about airport, aviation, and security in general. The threat is expanding interest in more reliable systems to detect presence of people within the protected perimeters. The occupancy sensors detect the presence of people (and sometimes animals) in a monitored area. Motion detectors respond only to moving objects. A distinction between the two is that the occupancy sensors produce signals whenever an object is stationary or not, while the motion detectors are selectively sensitive to moving objects. The applications of these sensors include security, surveillance, energy management (electric lights control), personal safety, friendly home appliances, point-of-sale advertisements, interactive toys, novelty products, etc. Depending on the applications, presence of humans may be detected through any means that is associated with some kind of a human body's property or body's actions [1]. For instance, a detector may be sensitive to body weight, heat, sounds, dielectric constant, etc. The following types of detectors are presently used for the occupancy and motion sensing of people:

1. *Air pressure sensors:* detect changes in air pressure resulted from opening doors and windows
2. *Capacitive:* detectors of human body capacitance
3. *Acoustic:* detectors of sound produced by people
4. *Photoelectric:* interruption of light beams by moving objects
5. *Optoelectric:* detection of variations in illumination or optical contrast in the protected area
6. *Pressure mat switches:* pressure sensitive long strips used on floors beneath the carpets to detect weight of an intruder
7. *Stress detectors:* strain gauges imbedded into floor beams, staircases, and other structural components
8. *Switch sensors:* electrical contacts connected to doors and windows

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_6, © Springer Science+Business Media, LLC 2010

9. *Magnetic switches:* a noncontact version of switch sensors
10. *Vibration detectors:* react to the vibration of walls or other building structures. Also, may be attached to doors or windows to detect movements
11. *Glass breakage detectors:* sensors reacting to specific vibrations produced by shattered glass
12. *Infrared motion detectors:* devices sensitive to heat waves emanated from warm or cold moving objects
13. *Microwave detectors:* active sensors responsive to microwave electromagnetic signals reflected from objects
14. *Ultrasonic detectors:* devices similar to microwave detectors except that instead of electromagnetic radiation, ultrasonic waves are used
15. *Video motion detectors:* a video equipment that compares a stationary image stored in memory with the current image from a protected area
16. *Video face recognition system:* image analyzers that compare facial features with database
17. *Laser system detectors:* similar to photoelectric detectors, except that they use narrow light beams and combinations of reflectors
18. *Triboelectric detectors:* sensors capable of detecting static electric charges carried by moving objects

One of the major aggravations in detecting occupancy or intrusion is a false-positive detection . The term "false-positive" means that the system indicates an intrusion when there is none. In some noncritical applications where false positive detections occur once in a while, for instance, in a toy or a motion switch controlling electric lights in a room, this may be not a serious problem: the lights will be erroneously turned on for a short time, which unlikely do any harm[1]. In other systems, especially used for the security and military purposes, false-positive detections, while generally not as dangerous as false negative ones (missing an intrusion), may become a serious problem[2]. While selecting a sensor for critical applications, considerations should be given to its reliability, selectivity, and noise immunity. It is often a good practice to form a multiple sensor arrangement with symmetrical interface circuits. It may dramatically improve a reliability of a system, especially in the presence of external transmitted noise. Another efficient way to reduce erroneous detections is to use sensors operating on different physical principles [2], for instance, combining capacitive and infrared detectors is an efficient combination as they are receptive to different kinds of transmitted noise.

[1]Perhaps just steering up some suspicion about living in a haunted house.

[2]A very nice movie "*How to Steal a Million*" (1966) was based on a plot where multiple false-positive alarms were so irritating that officials disabled the electronic protection system, exactly what the perpetrator was looking for.

6.1 Ultrasonic Detectors

These detectors are based on transmission to the object and receiving the reflected acoustic waves. A description of the ultrasonic detectors can be found in Sect. 7.5. For motion detection, they may require a somewhat longer operating range and a wider angle of coverage.

6.2 Microwave Motion Detectors

The microwave detectors offer an attractive alternative to other detectors when it is required to cover large areas and to operate over an extended temperature range under the influence of strong interferences, such as wind, acoustic noise, fog, dust, moisture, and so forth. These detectors (sensors) belong to the active sensors as they provide an excitation signal. That is, they emit pulses of the electromagnetic energy. Thus they can operate at day or night and do not rely on the external sources of energy. The operating principle of a microwave detector is based on radiation of electromagnetic radiofrequency (RF) waves toward a protected area. The electromagnetic waves backscattered (reflected) from objects whose sized are comparable with or larger than the wavelength of the excitation signal. The reflected waves are received, amplified, and analyzed. A time delay between the sent (pilot) signal and received reflected signal is used to measure distance to the object, while the frequency shift is used to measure speed of motion of the object.

The microwave detectors belong to the class of devices known as *radars*. Radar is an acronym for *RA*dio *D*etection *A*nd *R*anging.

The radar frequencies are as follows:

Band	Frequency range (GHz), f	Wavelength range (cm), λ
K_a	26.0–40.0	0.8–1.1
K	18.0–26.5	1.1–1.67
X	8.0–12.5	2.4–3.75
C	4.0–8.0	3.75–7.50
S	2.0–4.0	7.5–15
L	1.0–2.0	15–30
P	0.3–1.0	30–100

The name *microwave* is arbitrarily assigned to the wavelengths shorter than 4 cm (K_a, K, and X bands). They are long enough ($\lambda = 3$ cm at X band) to pass freely through most contaminants, such as clouds and airborne dust, and short enough for being reflected by larger objects. Other frequencies and types of energy (e.g., ultrasonic) are also used in a similar manner to the microwaves. For example, traffic controls use laser guns to identify speeders. In the laser detectors, series of short (nanosecond) pulses of infrared laser light is emitted, and the laser gun

measures the time it takes for the laser light to be reflected, just as in the microwave radars.

The microwave part of the detector consists of a Gunn oscillator, an antenna, and a mixer diode. The Gunn oscillator is a diode mounted in a small precision cavity, which, upon application of power, oscillates at microwave frequencies. The oscillator produces electromagnetic waves (frequency f_0), part of which is directed through an iris into a waveguide and focusing antenna that directs the radiation toward the object. Focusing characteristics of the antenna are determined by the application. As a general rule, the narrower the directional diagram of the antenna, the more sensitive it is (the antenna has a higher gain). Another general rule is that a narrow-beam antenna is much larger, whereas a wide-angle antenna can be quite small. The typical radiated power of the transmitter is 10–20 mW. A Gunn oscillator is sensitive to the stability of applied dc voltage and, therefore, must be powered by a good quality voltage regulator. The oscillator may run continuously, or it can be pulsed, which reduces the power consumption from the power supply.

A smaller part of the microwave oscillations is coupled to the Schottky mixing diode and serves as a reference signal (Fig. 6.1a). In many cases, the transmitter and the receiver are contained in one module called a transceiver. The target reflects some waves back toward the antenna, which directs the received radiation toward the mixing diode whose current contains a harmonic with a phase differential between the transmitted and reflected waves. The phase difference is in a direct relationship to the distance to the target. The phase-sensitive detector is useful mostly for detecting the distance to an object. However, in a motion sensor, movement, not distance, should be detected. Thus, for the occupancy and motion detector, the Doppler effect is the basis for the operation of microwave and ultrasonic detectors. It should be noted that the Doppler effect device is a true motion detector because it is responsive only to moving targets. Here is how it works.

An antenna transmits the frequency f_0, which is defined by the wavelength λ_0 as

$$f_0 = \frac{c_0}{\lambda_0}, \tag{6.1}$$

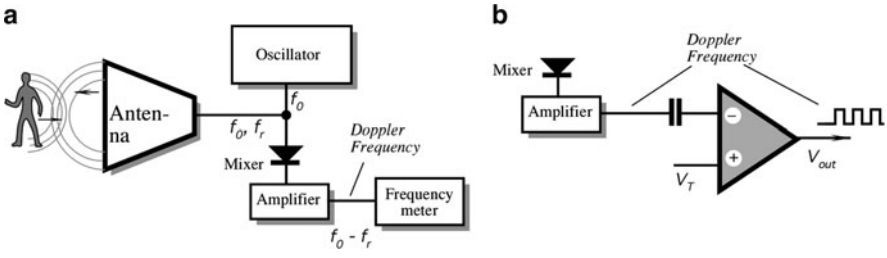

Fig. 6.1 Microwave occupancy detector: a circuit for measuring Doppler frequency (**a**); circuit with a threshold detector (**b**)

where c_0 is the speed of light in air. When the target moves toward or away from the transmitting antenna, the frequency of the reflected radiation will change. Thus, if the target is moving away with velocity v, the reflected frequency will decrease and it will increase for the approaching targets. This is called the Doppler effect, after the Austrian scientist Christian Johann Doppler (1803–1853)[3]. Although the effect first was discovered for sound, it is applicable to electromagnetic radiation as well. However, in contrast to sound waves that may propagate with the velocities dependent on movement of the source of sound, electromagnetic waves propagate with the speed of light, which is an absolute constant, independent of the light source movement. The frequency of reflected electromagnetic waves can be predicted by the Einstein's special theory of relativity as

$$f_{\mathrm{r}} = f_0 \frac{\sqrt{1 - \left(\frac{v}{c_0}\right)^2}}{1 + \frac{v}{c_0}}. \tag{6.2}$$

For practical purposes, when detecting a relatively slow-moving objects, the quantity $(v/c_0)^2$ is very small as compared with unity; hence, it can be ignored. Then, the equation for the frequency of the reflected waves becomes identical to that for the acoustic waves:

$$f_{\mathrm{r}} = f_0 \frac{1}{1 + \frac{v}{c_0}}. \tag{6.3}$$

As follows from (6.3), due to a Doppler effect, the reflected waves have a different frequency f_{r}. The mixing diode combines the radiated (reference) and reflected frequencies and, being a nonlinear device, produces a signal that contains multiple harmonics of both frequencies. The electric current through the diode may be represented by a polynomial:

$$i = i_0 + \sum_{k=1}^{n} a_k (U_1 \cos 2\pi f_0 t + U_2 \cos 2\pi f_r t)^k, \tag{6.4}$$

where i_0 is a dc component, a_k are the harmonic coefficients, which depend on a diode operating point, U_1 and U_2 are amplitudes of the reference and received signals, respectively, and t is time. The current through a diode contains an infinite number of harmonics, among which there is a harmonic of a differential frequency: $\Delta f = a_2 U_1 U_2 \cos 2\pi (f_0 - f_{\mathrm{r}})t$, which is called a Doppler frequency.

[3]During Doppler times, the acoustical instruments for precision measurements were not available yet. To prove his theory, Doppler placed trumpeters on a railroad flatcar and musicians with a sense of absolute pitch near the tracks. A locomotive engine pulled the flatcar back and forth at different speeds for two days. The musicians on the ground "recorded" the trumpet notes as the train approached and receded. The equations held up.

The Doppler frequency in the mixer can be found from (6.3):

$$\Delta f = f_0 - f_r = f_0 \frac{1}{1 + \frac{c_0}{v}}, \tag{6.5}$$

and since $c_0/v \gg 1$, the following holds after substituting (6.1):

$$\Delta f \approx \frac{v}{\lambda_0}. \tag{6.6}$$

Therefore, the signal frequency at the output of the mixer is linearly proportional to the velocity of a moving target. For instance, if a person walks toward the detectors with a velocity of 0.6 m/s, a Doppler frequency for the X-band detector is $\Delta f = 0.6/0.03 = 20$ Hz.

Equation (6.6) holds true only for movements in the normal direction. When the target moves at angles Θ with respect to the detector, the Doppler frequency is

$$\Delta f \approx \frac{v}{\lambda_0} \cos \Theta. \tag{6.7}$$

This implies that Doppler detectors theoretically become insensitive when a target moves at angles approaching 90°. In the velocity meters, to determine the velocity of a target, it is required is to measure the Doppler frequency and phase to determine direction of the movement (Fig. 6.1a). This method is used in police radars. For the supermarket door openers and security alarms, instead of measuring frequency, a threshold comparator is used to indicate the presence of a moving target (Fig. 6.1b). It should be noted that even if (6.7) predicts that the Doppler frequency is near zero for targets moving at angles $\Theta = 90°$, the entering of a target into a protected area at any angle results in an abrupt change in the received signal amplitude, and the output voltage from the mixer changes accordingly. Usually, this is sufficient to trigger the response of a threshold detector.

A signal from the mixer is in the range from microvolts to millivolts, so amplification is needed for the signal processing. Because the Doppler frequency is in the audio range, the amplifier is relatively simple. However, it generally must be accompanied by so-called notch filters, which reject a power line frequency and the main harmonic from full-wave rectifiers and fluorescent light fixtures: 60 and 120 Hz (or 50 and 100 Hz). For the normal operation, the received power must be sufficiently high. It depends on several factors, including the antenna aperture area A, target area a, and distance to the target r:

$$P_r = \rho \frac{P_0 A^2 a}{4\pi \lambda^2 r^4}, \tag{6.8}$$

where P_0 is the transmitted power. For effective operation, the target's cross-sectional area a must be relatively large, because for $\lambda^2 \leq a$, the received signal

is drastically reduced. Further, the reflectivity ρ of a target in the operating wavelength is also very important for the magnitude of the received signal. Generally, conductive materials and objects with high dielectric constants are good reflectors of electromagnetic radiation, whereas many dielectrics absorb energy and reflect very little. Plastics and ceramics are quite transmissive and can be used as windows in the microwave detectors.

The best target for a microwave detector is a smooth, flat conductive plate positioned normally toward the detector. A flat conductive surface makes a very good reflector; however, it may render the detector inoperable at angles other than 0°. Thus, an angle of $\Theta = 45°$ can completely divert a reflective signal from the receiving antenna. This method of diversion, along with other techniques, was used quite effectively in designs of the Stealth bombers, which are virtually invisible on radar screens.

To detect whether a target moves toward or away from the antenna, the Doppler concept can be extended by adding another mixing diode to the transceiver module. The second diode is located in the waveguide in such a manner that the Doppler signals from both diodes differ in phase by one-quarter of wavelength or by 90° (Fig. 6.2a). These outputs are amplified separately and converted into square pulses, which can be analyzed by a logic circuit. The circuit is a digital phase discriminator that determines the direction of motion (Fig. 6.2b). Door openers and traffic control are two major applications for this type of module. Both applications need the ability to acquire a great deal of information about the target for discrimination before enabling a response. In door openers, limiting the field of view and transmitted power may substantially reduce the number of false-positive detections. Although for door openers a direction discrimination is optional; for traffic control, it is a necessity to reject signals from the vehicles moving away. If the module is used for intrusion detection, the vibration of building structures may cause a large number of false-positive detections. A direction discriminator will respond to vibration with an alternate signal, and it will respond to an intruder with a steady

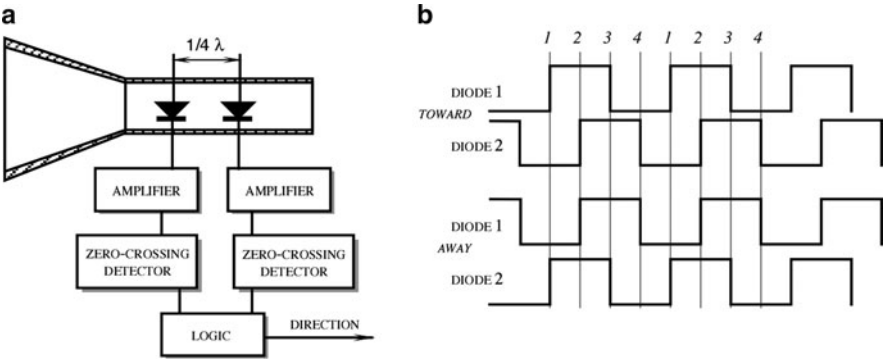

Fig. 6.2 Block diagram (**a**) and timing diagrams (**b**) of a microwave Doppler motion detector with directional sensitivity

logic signal. Hence, the direction discriminator is an efficient way to improve reliability of the detection.

Generally, the transmission and reception are alternated in time. That is, the receiver is disabled during the transmission; otherwise, a strong transmitted energy not only will saturate the receiving circuitry but may damage its sensitive components. In Nature, bats use ultrasonic ranging to catch their small prey. The bats become deaf for the short time when the ultrasonic burst of energy is transmitted. This temporary blinding of the receiver is the main reason why radars and acoustic rangers are not effective for short distances; it is just not enough time to disable and enable the receiver.

Whenever a microwave detector is used in the United States, it must comply with the strict requirements (e.g., MSM20100) imposed by the Federal Communication Commission. Similar regulations are enforced in many other countries. Also, emission of the transmitter must be below 10 mW/cm^2 as averaged over any 0.1-h period, as specified by OSHA 1910.97 for the frequency range from 100 MHz to 100 GHz.

A quite effective motion detector may be designed by employing micropower impulse radar (see Sect. 7.6.1). Advantages of such detectors are very low power consumption and nearly total invisibility to the intruder. The radar may be concealed inside wooden or masonic structures and is virtually undetectable by electronic means thanks to its low emission resembling natural thermal noise.

6.3 Capacitive Occupancy Detectors

Being a conductive medium with a high dielectric constant, a human body develops a coupling capacitance to its surroundings.[4] This capacitance greatly depends on such factors as body size, clothing, materials, type of surrounding objects, weather, and so forth. However wide the coupling range is, the capacitance may vary from few picofarads to several nanofarads. When a person moves, the coupling capacitance changes, thus making it possible to discriminate static objects from moving objects.

All objects form some degree of a capacitive coupling with respect to one another. If a human (or for that purpose – anything) moves into vicinity of the objects whose coupling capacitance with each other has been previously established, a new capacitive value arises between the objects as a result of presence of an intruding body. Figure 6.3 shows that the capacitance between a test plate and earth[5] is equal to C_1. When a person moves into vicinity of the plate, it forms two

[4]At 40 MHz, the dielectric constant of muscle, skin, and blood is about 97. For fat and bone, it is near 15.

[5]Here, by "earth" we mean any large object, such as the earth, lake, metal fence, car, ship, airplane, and so forth.

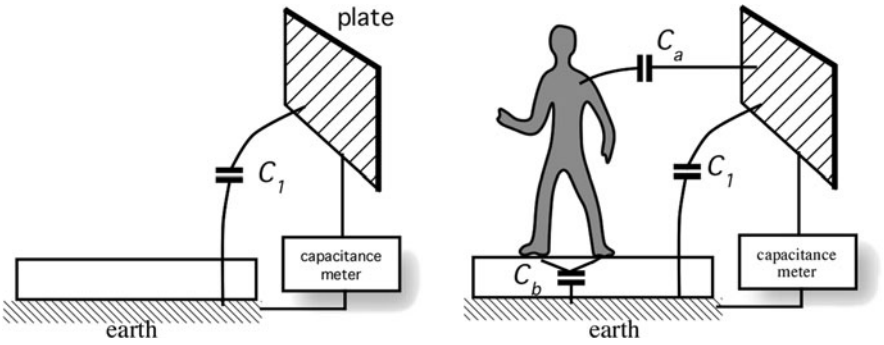

Fig. 6.3 An intruder brings in an additional capacitance to a detection circuit

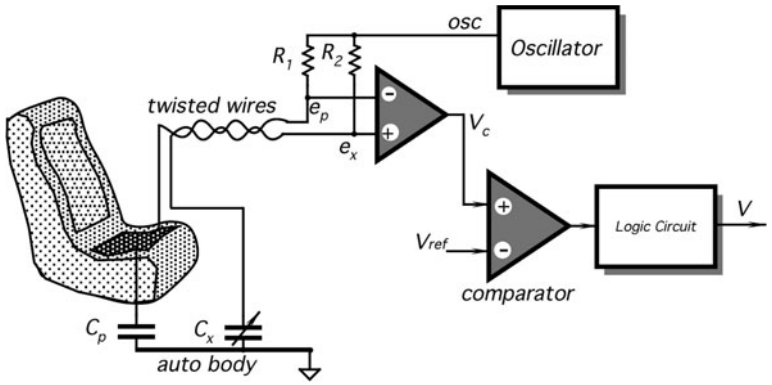

Fig. 6.4 Capacitive intrusion detector for automotive applications

additional capacitors: one between the plate and its own body, C_a, and the other between the body and the earth, C_b. Then, the resulting capacitance C between the plate and the earth becomes larger by the incremental capacitance ΔC.

$$C = C_1 + \Delta C = C_1 + \frac{C_a C_b}{C_a + C_b}. \tag{6.9}$$

With the appropriate apparatus, this phenomenon can be used for occupancy detection. What required is to measure capacitance between a test plate (the probe) and a reference plate (the earth).

Figure 6.4 illustrates a capacitive security system for an automobile [3]. A sensing probe is imbedded into a car seat. It can be fabricated as a metal plate, metal net, a conductive fabric, etc. The probe forms one plate of a capacitor C_p. The other plate of the capacitor is formed either by a body of an automobile, or by a separate plate positioned under a floor mat. A reference capacitor C_x is composed of

a simple fixed or trimming capacitor, which should be placed close to the seat probe. The probe plate and the reference capacitor are, respectively, connected to two inputs of a charge detector (resistors R_1 and R_2). The conductors preferably should be twisted to reduce the introduction of spurious signals as much as possible. For instance, strips of a twinflex cabling were found quite adequate. A differential charge detector is controlled by an oscillator, which produces square pulses (Fig. 6.5). Under a no-seat-occupied condition, the reference capacitor is adjusted to be approximately equal to C_p. Resistors and the corresponding capacitors define time constants of the networks. Both *RC* circuits have nearly equal time constants τ_1. Voltages across the resistors are fed into the inputs of a differential amplifier, whose output voltage V_c is near zero. Small spikes at the output is the result of some unavoidable imbalance. When a person is positioned on the seat, her body forms an additional capacitance in parallel with C_p, thus increasing a time constant of the R_1C_p-network from τ_1 to τ_2. This is indicated by the increased spike amplitudes at the output of a differential amplifier. The comparator compares V_c with a predetermined threshold voltage V_{ref}. When the spikes exceed the threshold, the comparator sends an indication signal to the logic circuit that generates signal V manifesting the car occupancy. It should be noted that a capacitive detector is an active sensor, because it essentially required an oscillating test signal to measure the capacitance value.

When a capacitive occupancy (proximity) sensor is used near or on a metal device, its sensitivity may be severely reduced due to a capacitive coupling between the electrode and the device's metallic parts. An effective way to reduce that stray capacitance is to use driven shields. Figure 6.6a shows a robot with a metal arm. The arm moves near people and other potentially conductive objects with which it could collide if the robot's control computer is not provided with an advance information on the proximity to the obstacles. An object, while approaching the arm, forms a capacitive coupling with it, which is equal to C_{so}. An arm is covered with an electrically isolated conductive sheath, which is called an *electrode*. As Fig. 6.3 shows, a coupling capacitance can be used to detect the proximity.

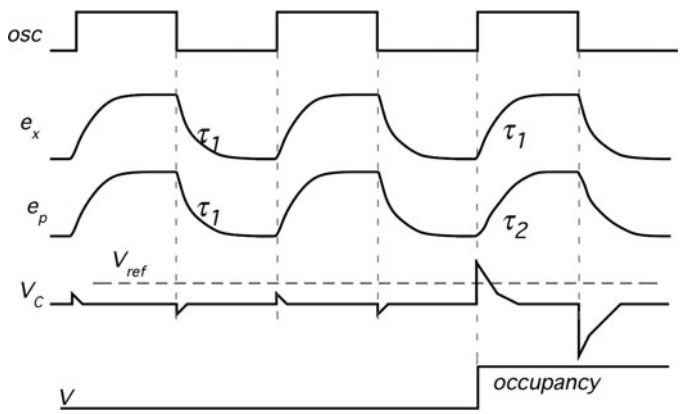

Fig. 6.5 Timing diagrams for a capacitive intrusion detector

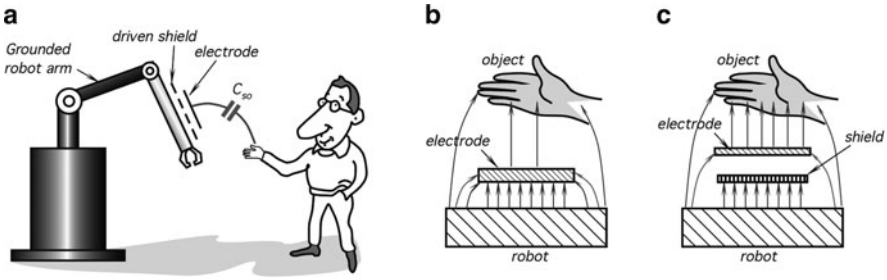

Fig. 6.6 Capacitive proximity sensor A driven shield is positioned on the metal arm of a grounded robot (**a**). Without the shield, the electric field is distributed between the electrode and the robot (**b**), while a driven shield directs electric field from the electrode toward the object (**c**)

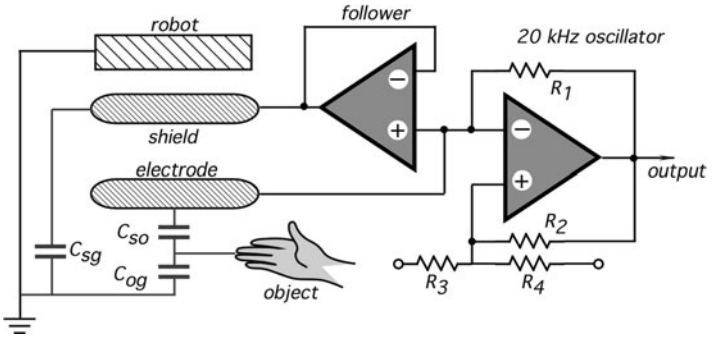

Fig. 6.7 Simplified circuit diagram of a frequency modulator controlled by the input capacitances

However, the nearby massive metal arm (Fig. 6.6b) forms a much stronger capacitive coupling with the electrode, which drugs the electric field from the object. An elegant solution[6] is to shield the electrode from the arm by an intermediate shield as shown in Fig. 6.6c. The sensor's assembly is a multilayer cover for the robotic arm, where the bottom layer is an insulator, then there is a large electrically conductive shield, then another layer of insulation, and on the top is a narrower sheet of the electrode. To reduce a capacitive coupling between the electrode and the arm, the shield must be at the same potential as the electrode, that is, its voltage must be driven by the electrode voltage (hence the name is a driven shield). As a result, no electric field is formed between them. Yet, a strong electric field will exist between the shield and the arm. The electric field is squeezed out from beneath the electrode and distributed toward the object.

Figure 6.7 shows a simplified circuit diagram of a square-wave oscillator whose frequency depends on the net input capacitance, comprised of C_{sg} (sensor-to-ground), C_{so} (sensor-to-object), and C_{og} (object-to-ground). The electrode is connected to the

[6]This device was developed for NASA's Jet Propulsion Laboratory by M.S. Katow at Planning Research Corp.

shield through a voltage follower. A frequency-modulated signal is fed into the robot's computer for controlling the arm movement. This arrangement allows to detect proximity to conductive objects over the range of 30 cm.

6.4 Triboelectric Detectors

Any object can accumulate, on its surface, static electricity. These naturally occurring charges arise from the triboelectric effect, which is a process of charge separation due to object movements, friction of clothing fibers, air turbulence, atmosphere electricity, etc. (see Sect. 3.1). Usually, air contains either positive or negative ions that can be attracted to the human body, thus modifying its charge. Under the idealized static conditions, an object is not charged; its bulk charge is equal to zero. In reality, any object, which at least temporarily is isolated from the ground, can exhibit some degree of its bulk charge imbalance. In other words, it becomes a carrier of electric charges.

Any electronic circuit is made of conductors and dielectrics. If a circuit is not shielded, all its components exhibit a certain capacitive coupling to the surrounding objects. A pick-up electrode can be added to the circuit's input to increase its coupling to the environment, very much like in the capacitive detectors that were covered in the previous Sect. 6.3. The electrode can be fabricated in form of a conductive surface that is well isolated from the ground. The difference between the triboelectric and capacitive sensors is that in the former no capacitance is being measured but rather an electric charge that is accumulated on the capacitance.

Electric field is established between the surrounding objects and the electrode whenever at least one of them carries electric charges. In other words, all distributed capacitors formed between the electrode and the environmental objects are charged by static or slow changing electric fields resulted from a triboelectric effect. Under a no-occupancy condition, electric field in the electrode vicinity is either constant or changes relatively slow.

If a charge carrier (a human or an animal) changes its position and moves away or a new charge carrying an object enters into the vicinity of the electrode, a static electric field is disturbed. This results in a redistribution of charges between the coupling capacitors, including those which are formed between the input electrode and the surroundings. The charge magnitude depends on the atmospheric conditions and nature of the objects. For instance, a person in dry man-made clothes[7] walking along a carpet carries a million times stronger charge than a wet intruder who has come from the rain. An electronic circuit can be adapted to sense these variable charges at its input. In other words, it can be made capable of converting the induced variable charges into electric signals that may be amplified and further processed. Thus, static electricity, which is a naturally occurring phenomenon, can

[7]Many "man-made" objects are made by women, so do not look for any sexism here.

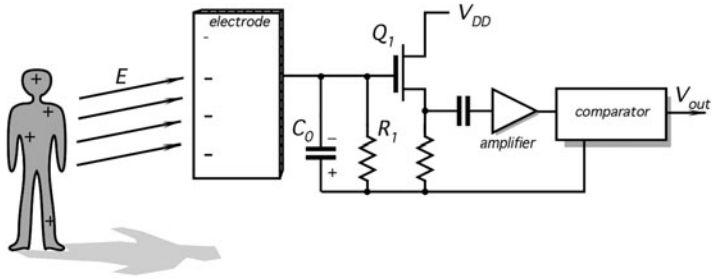

Fig. 6.8 Monopolar triboelectric motion detector

be utilized to generate alternating signals in the electronic circuit to indicate the movement of objects.

Figure 6.8 shows a monopolar triboelectric motion detector. It is comprised of a conductive electrode connected to an analog impedance converter made with a MOS transistor Q_1, a bias resistor R_1, an input capacitance C_o, a gain stage, and a window comparator [4]. While the rest of the electronic circuit may be shielded, the sensing electrode is exposed to the environment and forms a coupling capacitor C_p with the surrounding objects. In Fig. 6.8, static electricity is exemplified by positive charges distributed along the person's body. Being a charge carrier, the person becomes a source of an electric field, having intensity E. The field induces a charge of the opposite sign in the electrode. Under the static conditions, when the person does not move, the field intensity is constant and the input capacitance C_o is discharged through a bias resistor R_1. To make the circuit sensitive to relatively slow motions, the resistor R_1 should be selected of a very high value: on the order of 1010 Ω or higher. When the person moves, intensity E of the electric field changes. This induces a corresponding variable electric charge in the input capacitor C_o and results in appearance of a variable electric voltage across the bias resistor. That voltage is fed through the coupling capacitor into the gain stage whose output signal is applied to a window comparator. The comparator compares the signal with two thresholds, as it is illustrated in a timing diagram of Fig. 6.9b. A positive threshold is normally higher than the baseline static signal, while the negative threshold is lower. During human movement, a signal at the comparator's input deflects either upward or downward, crossing one of the thresholds. Output signals from the window comparator are square pulses which can be utilized and further processed by the conventional data processing devices. It should be noted that contrary to a capacitive motion detector, which is an active sensor, a triboelectric detector is passive. That is, it does not generate or transmit any signal that makes it more difficult to detect. This detector may be hidden in or behind nonmetallic objects such as wood, bricks, etc.

There are several possible sources of interferences that may cause spurious detections by the triboelectric detectors. The detector may be subjected to a transmitted noise resulting in false-positive detection. Among the noise sources are 60 or 50 Hz power line signals, electromagnetic fields generated by radio

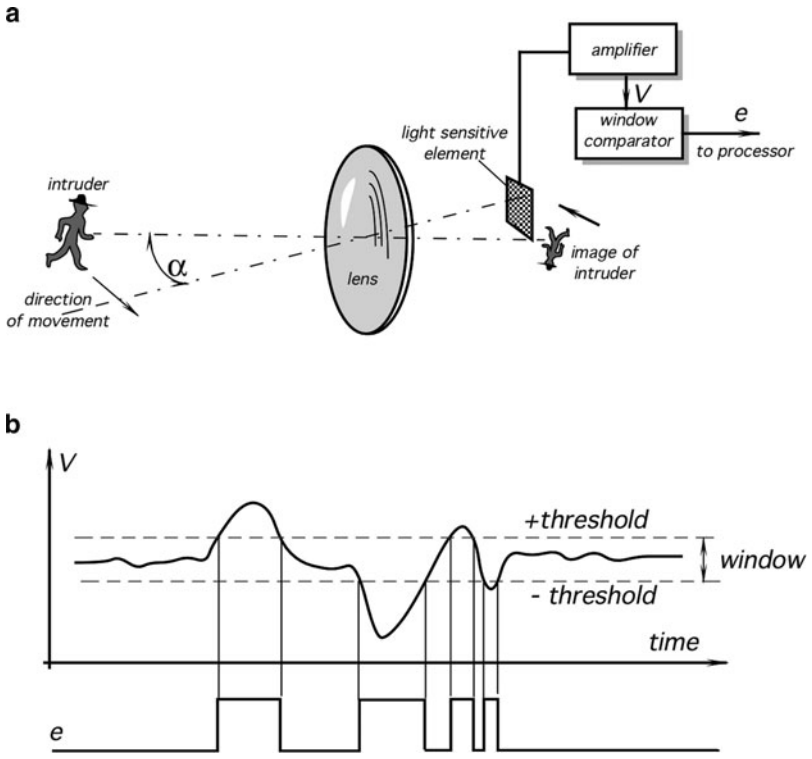

Fig. 6.9 General arrangement of an optoelectronic motion detector. A lens forms an image of a moving object (intruder). When the image crosses the sensor's optical axis it covers the sensitive element (**a**). The element responds with a signal that is amplified and compared with a window threshold in a comparator (**b**)

stations, power electric equipment, lightnings, etc. Most of these interferences generate electric fields, which are distributed around the detector quite uniformly and can be compensated for by employing a symmetrical input circuit with a significant common mode rejection ratio.

6.5 Optoelectronic Motion Detectors

By far the most popular intrusion sensors are the optoelectronic motion detectors. They rely on electromagnetic radiation in the optical range, specifically having wavelengths from 0.4 to 20 μm. This covers the visible, near, and part of far infrared (IR) spectral ranges. The detectors are primarily used for indication of movement of people and animals. They operate over distances ranging up to several hundred meters and, depending on the particular need, may have either a narrow or wide field of view.

The operating principle of the optical motion detectors is based on detection of light (either visible or infrared) reflected or emanated from surface of a moving object into the surrounding space. Such radiation may be originated either by an external light source and then reflected by the object or it may be produced by the object itself in form of a natural IR emission. The former case is classified as an active detector and the latter a passive detector. Hence, an active detector requires an additional light source, for instance, daylight, electric lamp, infrared (IR) light emitting diode (LED) projector, laser, etc. The passive detectors perceive mid- and far-infrared natural emission from objects having temperatures that are different from the surroundings. Both types of detectors use an optical contrast as means of the object recognition.

First, we shall consider limitations of the optoelectronic detectors as opposed to such devices as microwave or ultrasonic devices. Presently, optoelectronic detectors are used almost exclusively to detect presence or absence of movement qualitatively rather than quantitatively. In other words, the optoelectronic detectors are very useful to indicate whether an object moves or not, while they can not distinguish one moving object from another and they cannot be utilized to accurately measure distance to a moving object or its velocity. The major application areas for the optoelectronic motion detectors are in security systems (to detect intruders), in energy management (to turn lights on and off), and in the so-called "smart" homes where they can control various appliances, such as air conditioners, cooling fans, stereo players, etc. They also may be used in robots, toys, point-of-sale advertisements, and novelty products. The most important advantage of an optoelectronic motion detector is simplicity and low cost.

6.5.1 Sensor Structures

A general structure of an optoelectronic motion detector is shown in Fig. 6.9a. Regardless what kind of sensing element is employed, the following components are essential: a focusing device (a lens or curved mirror), a light detecting element, and a threshold comparator. An optoelectronic motion detector resembles a photographic camera. Its focusing component creates on a focal plane an image of the field of view. While there is no shutter like in a camera, in place of an imaging sensor, a light sensitive element is used. The element converts the focused light into an electric signal. Since no real image needs to be processed, such an element can be considered as a single-pixel opto-electronic detector[8].

Let us assume that the motion detector is mounted in a room. A focusing lens creates an image of the room on a focal plane where the light-sensitive element is positioned. If the room is unoccupied, the image is static and the output signal from the element is steady stable. When an "intruder" enters the room and keeps moving,

[8]In a differential sensor, as described below, two "pixels" are employed.

her image on the focal plane also moves. In a certain moment, the intruder's body is displaced for an angle α and the image overlaps with the element. This is an important point to understand: the detection is produced only at the moment when the object's image either coincides with the detector's surface or clears it. That is, no crossing, no detection. Assuming that the intruder's body creates an image whose photon flux is different from that of the static surroundings, the light-sensitive element responds with a deflecting voltage V. In other words, to cause detection, a moving image shall have a certain degree of an optical contrast with its surroundings.

Figure 6.9b shows that the output signal is compared with two thresholds in a window comparator. The purpose of the comparator is to convert the analog signal V into two logic levels: 0, no motion detected and 1, motion is detected. In most cases, signal V from the element first must be amplified and conditioned before it becomes suitable for the threshold comparison. The window comparator contains both the positive and negative thresholds, while signal V is positioned in-between. Whenever image of a moving object overlaps with the light-sensitive element, voltage V deflects from its steady-state position and crosses one of two thresholds. The comparator generates a positive voltage (1), thus indicating detection of movement in the field of view. The operation of this circuit is identical to the threshold circuits described earlier for other types of occupancy detectors.

It may be noted from Fig. 6.9 that the detector has quite a narrow field of view : if the intruder keeps moving, her image will overlap with the sensor only once. After that the window comparator output will produce steady zero. This is the result of a small area of the sensing element. In some instances, when a narrow field of view is required it is quite all right, however, in the majority of cases, a much wider field of view is desirable. This can be achieved by several methods described below.

6.5.1.1 Multiple Detecting Elements

An array of detecting elements (multiple pixels) may be placed in the focal plane of a focusing mirror or lens. Each individual element covers a narrow field of view, while in combination they protect larger area. All detectors in the array either shall be multiplexed or otherwise interconnected to produce a combined detection signal.

6.5.1.2 Complex Sensor Shape

If the detecting element's surface area is sufficiently large to cover the entire angle of view, the area may be optically broken into smaller elements, thus creating an equivalent of a multiple detector array. To break the surface area into several parts, one may shape the sensing (detecting) element in an odd pattern like that shown in Fig. 6.10a. Each part of the element acts as a separate light detector. All such detectors are electrically connected either in parallel or in series, being arranged in a serpentine pattern. The parallel or serially connected detectors generate a combined

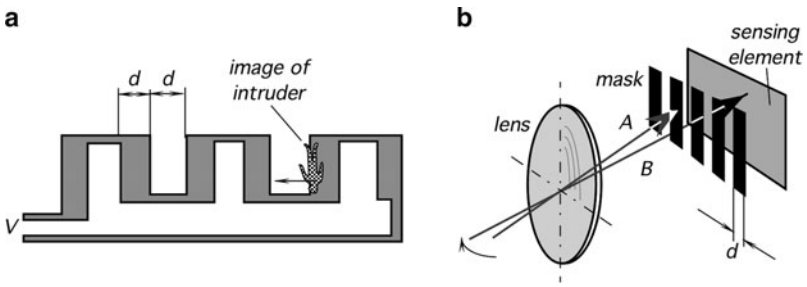

Fig. 6.10 Complex shape of a sensing element (**a**); and image distortion mask (**b**)

output signal, for instance, voltage v, when image of the object moves along the element surface crossing alternatively sensitive and nonsensitive areas. This results in an alternate signal v at the detector terminals. For a better sensitivity, such sensitive and nonsensitive areas should be sufficiently large to overlap with most of the object's image.

6.5.1.3 Image Distortion

Instead of making the detector in a complex shape, an image of an entire filed of view may be broken into several parts. This can be done by placing a distortion mask [5] in front of the detector having a sufficiently large area as it is depicted in Fig. 6.10b. The mask is opaque and allows formation of an image on the detector's surface only within its clearings. The mask operation is analogous to the complex sensor's shape as described above.

6.5.1.4 Facet Focusing Element

Another way of broadening the field of view while employing a small area detector is to use multiple focusing devices. A focusing mirror or a lens may be divided into arrays of smaller mirrors or lenses called *facets*, resembling an eye of an insect. Each facet works as an individual lens (mirror) creating its own image on a common focal plane. All facets form multiple images as shown in Fig. 6.11a. When the object moves, the images also move across the sensing element, causing it to produce an alternate signal. By combining multiple facets, it is possible to shape any desirable detecting pattern in the field of view, in both horizontal and vertical planes. Positioning of the facet lens, focal distances, number, and a pitch of the facets (a distance between the optical axes of two adjacent facets) may by calculated in every case by applying rules of geometrical optics. The following practical formulas may be applied to find the focal length of a facet

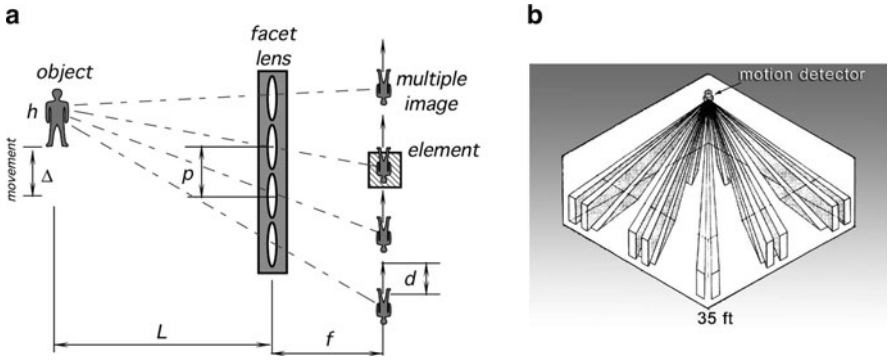

Fig. 6.11 A facet lens creates multiple images near the sensing element (**a**); Sensitive zones created by a complex facet lens (**b**)

$$f = \frac{Ld}{\Delta}, \tag{6.10}$$

and the facet pitch is

$$p = 2nd, \tag{6.11}$$

where L is distance to the object, d is width of the sensing element, n is number of the sensing elements (evenly spaced), and Δ is the object's minimum displacement, which must result in detection.

For example, if the sensor has two sensing elements of $d = 1$ mm each, which are positioned at 1 mm apart, and the object's minimum displacement $\Delta = 25$ cm at a distance $L = 10$ m, the facet focal length is calculated from (6.10) as $f = (1{,}000\text{ cm})(0.1\text{ cm})/25\text{ cm} = 4$ cm, and the facets should be positioned with a pitch of $p = 8$ mm from one another as per (6.11).

By combining facets, one may design a lens that covers a large field of view (Fig. 6.11b) where each facet creates a relatively narrow angle-sensitive zone. Each zone projects an image of an object into the same sensing element. When the object moves, it crosses the zone boundaries, thus modulating the sensor's output. Currently, the facet lenses are primarily used in the mid- and far-infrared spectral ranges. These lenses are molded of high-density polyethylene (HDP) (Fig. 6.12) and quite inexpensive.

6.5.2 *Visible and Near IR Light Motion Detectors*

Most of objects (apart from very hot) radiate electromagnetic waves only in the mid- and far-infrared spectral ranges. Hence, visible and near infrared light motion detectors have to rely on additional sources of light that illuminate objects. The

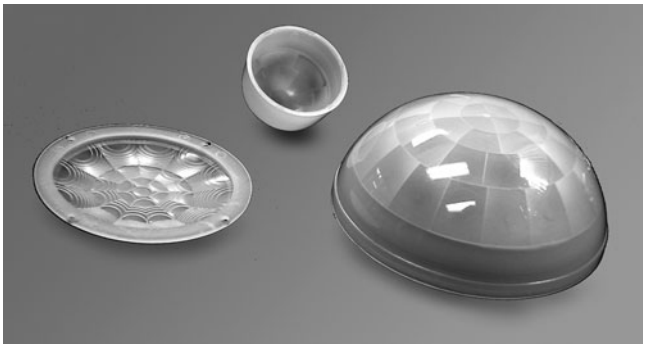

Fig. 6.12 Examples of three-dimensional infrared faceted Fresnel lenses molded of HDP

light is reflected by the object's surface toward the focusing device for a subsequent detection. The light sources may be sun, incandescent lamps, or the invisible infrared (IR) LEDs. The use of visible light for detecting moving objects goes back to 1932 [5] when in the preradar era, inventors were looking for ways to detect moving cars and flying airplanes. In one invention, an airplane detector was built in form of a photographic camera where the focusing lens made of glass was aimed at the sky. A moving plane's image was focused on a selenium photodetector, which reacted to a changing contrast in the sky image. Naturally, such a detector could operate only at a daytime to detect planes flying below clouds. Obviously, those detectors were not too practical. Another version of a visible range light motion detector was patented and produced for less demanding applications: controlling lights in a room [4] and to make interactive toys [6].

To turn the lights off in a nonoccupied room, the visible range motion detector[9] was combined with a timer and a solid-state relay. The detector was activated when the room is illuminated. Visible light photons carry a relatively high energy and may be detected by quantum photovoltaic or photoconductive cells whose detectivity is quite high (see Chap. 14). Thus, the optical system may be substantially simplified. In the motion switch, the focusing device was built in form of a pinhole lens (Figs. 6.13b and 6.14a). Such a lens is just a tiny hole in an opaque foil. To avoid a light diffraction, a hole diameter must be substantially larger than the longest detectable wavelength (red). The motion switch had a three-facet pinhole lens with apertures of 0.2 mm in diameter (Fig. 6.13c). A pinhole lens has a theoretically infinite depth of focusing range; hence, a photodetector can be positioned at any distance from the lens. For practical reasons, this distance was calculated for the object's displacement, view angle, and the photoresistor dimensions used in the design. The photoresistor was selected with a serpentine pattern of a sensing element (Fig. 6.13a) and connected into a low frequency a.c. amplifier. When the room was illuminated, the motion sensor acted as a miniature photographic camera:

[9]"Motion Switch" of Fig. 6.14a for some time was manufactured by Intermatic, Inc.

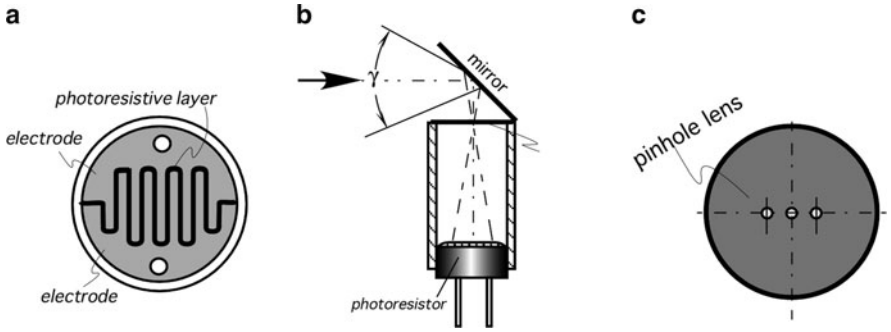

Fig. 6.13 A simple optical motion detector for a light switch and toys: (**a**) a sensitive surface of a photoresistor forms a complex sensing element; (**b**) a flat mirror and a pinhole lens form an image on a surface of the photoresistor; (**c**) pinhole lenses

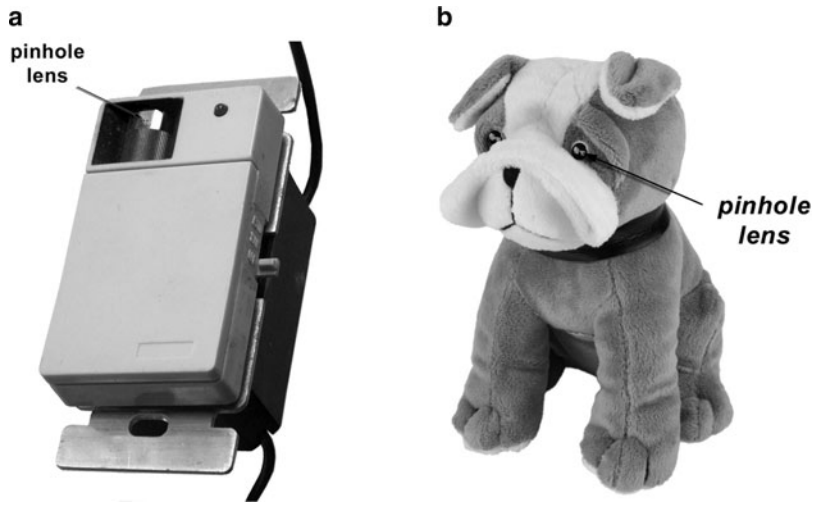

Fig. 6.14 Motion sensing light switch with a photoresistor and pinhole lens (**a**), interactive toy (**b**) that reacts to a child movement; the dog barks when motion is detected

an image from the lens' field of view was created on a surface of the photoresistor. Moving people in the room caused the image to change in such a way as the optical contrast changed across the serpentine pattern of the photoresistor. In turn, its resistive value was changing, resulting in modulation of the electric current passing through the photoresistor. This signal was further amplified and compared with a predetermined threshold. Upon crossing that threshold, the comparator generated electric pulses that reset a 15-min timer. If no motion was detected within 15 min from the last movement, the timer disabled the solid-state relay to turn lights off. Then, light could be turned on back only manually, because this motion detector cannot function in darkness. Thanks to its low cost, this type of a motion sensor was

used in interactive toys that react to movement of children [6]. An example of such a toy is shown in Fig. 6.14b where a pinhole lens was built into an eye of a mechanized barking dog. Normally the dog was sitting quietly but when motion in its vicinity was detected, the dog started moving and barking. If you pet it on the back, the barking stopped and the dog wagged the tail (a tactile sensor was installed on the back under the coat (see Sect. 9.2).

6.5.3 Far-Infrared Motion Detectors

Another version of a motion detector operates in the optical range of thermal radiation, the other name for which is mid- and far-infrared (IR). Such detectors are responsive to radiative heat exchange between the sensing element and moving object [7–9]. Here we will discuss a detection of moving people, however the technique which is described below is applicable for other warm or cold objects.

The principle of thermal motion detection is based on the physical theory of natural emission of electromagnetic radiation from any object whose temperature is above absolute zero. The fundamentals of this theory are described in Sect. 3.12.3. We recommend that the reader first familiarize herself with that section before going further.

For motion detection, it is essential that surface temperature of an object be different from that of the surrounding objects, so a thermal contrast would exist, just as a visible contrast in the sensors described above. All objects emanate thermal radiation from their surfaces and intensity of that radiation is governed by the Stefan-Boltzmann law (3.133). If the object is warmer than the surroundings, its thermal radiation is shifted toward shorter wavelengths and the intensity becomes stronger. Most objects whose movement is to be detected are nonmetals, hence they radiate thermal energy quite uniformly within a hemisphere (Fig. 3.45a). Moreover, the dielectric objects generally have high emissivity of thermal radiation. Human skin is a quite good emitter of thermal radiation. Its emissivity is well over 90% (see A.18). Most of natural and synthetic fabrics also have high emissivities between 0.74 and 0.95. There are two types of the IR motion detectors: passive and active.

6.5.3.1 Passive Infrared Motion Detectors

The passive infrared (PIR) motion detectors became very popular for the security and energy management systems. The PIR sensing element is responsive to mid- and far-infrared radiation within a spectral range from approximately 4 to 20 μm where most of the thermal power emanated by humans is concentrated (surface temperatures ranging from about 28 to 37°). There are three types of sensing elements that are potentially useful for that detector: bolometers, thermopiles, and pyroelectrics; however, the pyroelectric elements are used almost exclusively for motion detection thanks to their simplicity, low cost, high responsivity, and a broad dynamic range. A pyroelectric effect is described in Sect. 3.7 and some detectors are

covered in Sect. 14.7.3. Here, we are going to see how that effect may be employed in a practical motion sensor design.

A pyroelectric material generates an electric charge in response to a thermal energy flow through its body. In a very simplified way it may be described as a secondary effect of a thermal expansion (Fig. 6.15). Since all pyroelectrics are also piezoelectrics, the absorbed heat causes the front side of the sensing element to expand. The resulting thermally induced stress leads to development of a piezoelectric charge on the element electrodes. This charge is manifested as voltage across the electrodes deposited on the opposite sides of the material. Unfortunately, the piezoelectric properties of the element have also a negative effect. If the sensor is subjected to a minute mechanical stress due to any external force, like sounds or structural vibrations, it also generates a charge, which in most cases is indistinguishable from that caused by the infrared heat waves.

To separate thermally induced charges from the piezoelectrically induced charges, a pyroelectric sensor is usually fabricated in a symmetrical form (Fig. 6.16a). Two identical elements are positioned inside the sensor's housing. The elements

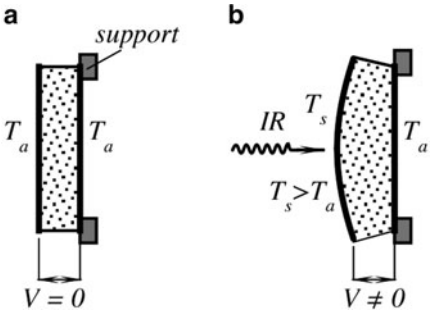

Fig. 6.15 Simplified model of a pyroelectric effect as a secondary effect of piezoelectricity. Initially, the element has a uniform temperature (**a**); upon exposure to thermal radiation, its front side warms up and expands, causing a stress-induced charge (**b**)

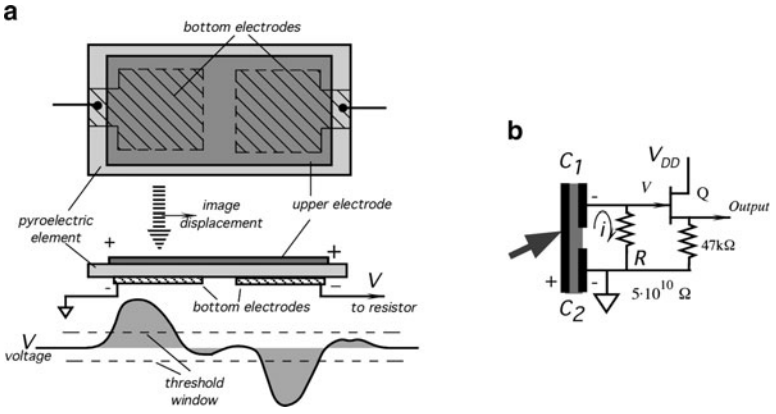

Fig. 6.16 Dual pyroelectric sensor
Sensing element with a front (*upper*) electrode and two bottom electrodes deposited on a common crystalline substrate (**a**). A moving thermal image travels from left part of the sensor to the right generating an alternate voltage across the bias resistor, *R* (**b**)

are connected to the electronic circuit in such a manner as to produce the out-of-phase signals when subjected to the same in-phase inputs. The idea is that interferences that are produced by the piezoelectric effect or spurious heat signals are applied to both electrodes simultaneously (in phase) and thus canceled at the input of the electronic circuit, while the variable thermal radiation will be absorbed by only one element at a time, thus avoiding a cancellation. This arrangement is called a differential sensor.

One way to fabricate a differential sensor is to deposit two pairs of electrodes on both sides of a pyroelectric element. Each pair forms a capacitor, which may be charged either by heat or by a mechanical stress. The electrodes on the upper side of the sensor are connected together forming one continuous electrode, while the two bottom electrodes are separated, thus creating the opposite-serially connected capacitors. Depending on the side where the electrodes are positioned, the output signal will have either a positive or negative polarity for a thermal influx. In same applications, a more complex pattern of the sensing electrodes may be required (for instance, to form predetermined detection zones), so that more than one pair of the electrodes are needed. In such a case, for better rejection of the in-phase signals (common mode rejection) the sensor still should have an even number of pairs where positions of the pairs alternate for a better geometrical symmetry. Sometimes, such an alternating connection is called an interdigitized electrode.

A symmetrical sensing element should be mounted in a way to assure that both parts of the element generate the same signal if subjected to the same external factors. At any moment, the optical component (e.g., a Fresnel lens) must focus a thermal image of an object on the surface of one part of the sensor only, otherwise signals from the image will be cancelled. The element generates a charge only across the electrode pair, which is subjected to a heat flux. When a thermal image moves from one electrode to another, the current i flowing from the sensing element to the bias resistor R (Fig. 6.16b) changes from zero, to positive, then back to zero, then to negative, and again back to zero (Fig. 6.16a, lower portion). A JFET transistor Q is used as an impedance converter. Resistor R value must be very high. For example, a typical alternate current generated by the element in response to a moving person is on the order of 1 pA (10^{-12} A). If a desirable output voltage for a specific distance is $v = 50$ mV, according to Ohm's law the resistor value should be $R = v/i = 50$ GΩ (5×10^{10} Ω). Such a resistor cannot be directly connected to a regular electronic circuit, hence transistor Q serves as a voltage follower (the gain is close to unity). Its typical output impedance is on the order of several kilohms.

Table A.9 lists several crystalline materials which possess pyroelectric properties and can be used for fabrication of sensing elements. Most often used are ceramic elements thanks to their low cost and ease of fabrication. A pyroelectric coefficient of ceramics to some degree may be controlled by varying their porosity (creating voids inside the sensor's body). An interesting pyroelectric material is a polymer film PVDF, which while being not as sensitive as most of the solid-state crystals, has advantages of being flexible and inexpensive. Besides, it can be produced in any size and may be bent or folded in any desirable fashion (see Sect. 3.6.2).

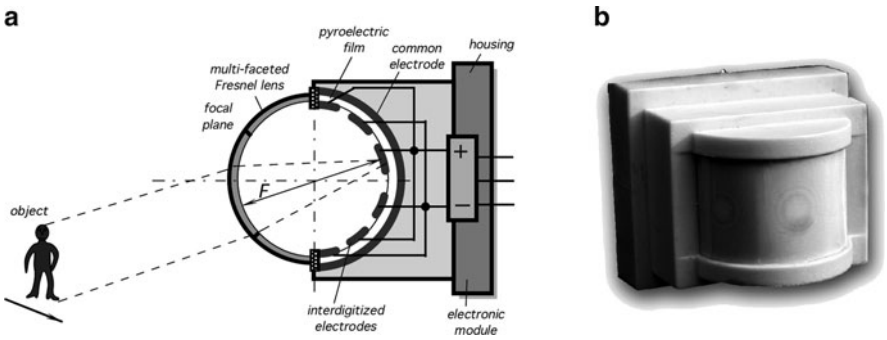

Fig. 6.17 Far infrared motion detector uses a curved Fresnel lens and a pyroelectric PVDF film. Internal structure of the sensor (**a**) and external appearance of the sensor (**b**)

Besides the sensing element, an infrared motion detector needs a focusing device. Some detectors employ parabolic mirrors while the Fresnel plastic lenses (Sect. 4.6) become more and more popular because they are inexpensive, may be molded in any desirable shape (Fig. 6.12), and in addition to focusing, act as windows to protect the interior of the sensor from outside moisture and pollutants.

To illustrate how a plastic Fresnel lens and a PVDF film can work together, let us look at the motion detector depicted in Fig. 6.17a. It uses a HDP multifaceted curved lens and a curved PVDF film sensor [7]. The sensor design combines two methods described above: a facet lens and a complex electrode shape. The lens and the film are curved with the same radii of curvature equal to one-half of the focal distance *f*, thus assuring that the film is always positioned in the focal plane of the corresponding facet of the lens. The film has a pair of large interdigitized electrodes, which are connected to the positive and negative input of a differential amplifier located in the electronic module. The amplifier rejects common-mode interference and amplifies a thermally induced voltage. The side of the film facing the lens is coated with an organic coating to improve its absorptivity in the far infrared spectral range. This design results in a fine resolution (detection of small displacement at a longer distance) and a small size of the sensor (Fig. 6.17b). Small sensors are especially useful for installation in devices where overall dimensions are critical. For instance, one such application is a light switch where the detector must be mounted into a wall plate of the switch.

6.5.3.2 PIR Sensor Efficiency Analysis

Regardless of the type of optical device employed, majority of modern PIR detectors operate on the same physical effect - pyroelectricity. To analyze performance of such a sensor, first we must calculate the infrared power (flux), which is converted into an electric charge by the sensing element. The optical device focuses thermal radiation, forming a miniature thermal image on the surface of the sensor.

The photon energy of an image is absorbed by the sensing element and is converted into heat by a heat-absorbing surface. That heat, in turn, is converted by the pyroelectric crystalline element into a minute electric charge. And finally, the charge causes a very small electric current passing through the input of an interface circuit. A maximum operating distance for given conditions can be determined by the noise level of the detector. For reliable discrimination, the worst-case noise power must be at least ten times smaller than that of the signal. The pyroelectric sensor is a converter of thermal energy flow into electric charge. The energy flow essentially demands a presence of a thermal gradient across the sensing element. In the detector, the element of thickness h has the front side exposed to the lens, while the opposite side faces the detector's interior housing, which normally is at ambient temperature T_a. The front side of the sensor element is covered with a heat absorbing coating to increase its emissivity ε_s to the highest possible level, preferably close to unity. When thermal flux Φ_s is absorbed by the element's front side, the temperature goes up and heat starts propagating through the sensor toward its rear side. Thanks to the pyroelectric properties, electric charge is developing on the element surfaces in response to the heat flow.

To estimate a power level at the sensor's heat absorbing surface, let us make some assumptions. We assume that the moving object is a person whose effective surface area is b (Fig. 6.18), the temperature along this surface (T_b) is distributed uniformly and is expressed in Kelvin. The object is moving at a distance L from the motion detector and being a diffuse emitter it radiates IR energy uniformly within the hemisphere having a surface area of $A = 2\pi L^2$. Also, we assume that the focusing device makes a sharp image of the object. For this calculation, we select a lens, which has a surface area a. The sensor's temperature in K is T_a, the same as that of ambient.

Total infrared power (flux) lost to surroundings from the object can be determined from the Stefan-Boltzmann law

$$\Phi = b\varepsilon_a\varepsilon_b\sigma\left(T_b^4 - T_a^4\right), \tag{6.12}$$

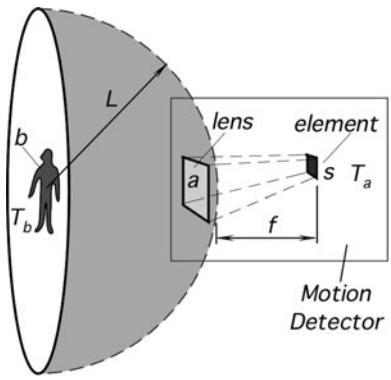

Fig. 6.18 Formation of a thermal image on the sensing element of a PIR motion detector

where σ is the Stefan-Boltzmann constant, ε_b and ε_a are the object and the surrounding emissivities, respectively. If the object is warmer than the surroundings (which is usually the case), the net infrared power is distributed toward an open space having ambient temperature T_a. Since the object is a diffusive emitter, we may consider that the same flux density may be detected at any point along the equidistant surface. In other words, the intensity of infrared power is distributed uniformly along the spherical surface having radius L. The above assumptions are rather stretched, yet they allow to estimate what portion of the IR flux is received by the lens.

Assuming that the surroundings and the object's surface are ideal emitters and absorbers ($\varepsilon_b = \varepsilon_a = 1$) and the sensing element's emissivity is ε_s, the net radiative flux density at distance L can be derived as

$$\phi = \frac{b}{2\pi L^2}\varepsilon_\mathrm{s}\sigma\left(T_b^4 - T_a^4\right). \tag{6.13}$$

The lens efficiency (transmission coefficient) is γ, which theoretically may vary from 0 to 0.92 depending on properties of the lens material and the lens design. For the HDP Fresnel lenses, the transmission value is in the range from 0.5 to 0.75. After ignoring a minor nonlinearity related to the fourth power of temperatures in (6.13), thermal power absorbed by the element can be expressed as

$$\Phi_\mathrm{s} \approx a\gamma\phi \approx \frac{2\sigma\varepsilon_s}{\pi L^2}a\gamma T_\mathrm{a}^3(T_b - T_a). \tag{6.14}$$

It is seen from (6.14) that the infrared flux, which is focused by the lens on the surface of the sensing element, is inversely proportional to the squared distance from the object and directly proportional to the areas of the lens and the object. It is important to note that in the case of a multifacet lens, the lens area a relates only to a single facet and not to a total lens area.

If the object is warmer than the sensor, the flux Φ_s is positive. If the object is cooler, the flux becomes negative, meaning it changes its direction: the heat goes from the sensor to the object. In reality, this may happen when a person walks into a warm room from the cold outside. The surface of her clothing will be cooler than the sensor and the flux will be negative. In the following discussion, we will consider that the object is warmer than the sensor and the flux is positive.

Upon influx of the infrared radiation, temperature of the sensor element increases (or decreases) with a rate that can be derived from the absorbed thermal power Φ_s and thermal capacity C of the element

$$\frac{dT}{dt} \approx \frac{\Phi_\mathrm{s}}{C}, \tag{6.15}$$

where t is time. This equation is valid during a relatively short interval, immediately after the sensor is exposed to the thermal flux, and can be used to evaluate the signal magnitude.

The electric current generated by the sensor can be found from the fundamental formula:

$$i = \frac{dQ}{dt}, \tag{6.16}$$

where Q is the electric charge developed by the pyroelectric sensor. This charge depends on the sensor's pyroelectric coefficient P, sensor's area s and temperature change dT:

$$dQ = PsdT. \tag{6.17}$$

Thermal capacity C can be derived through a specific heat c of the material, area s, and thickness of the element h:

$$C = csh. \tag{6.18}$$

By substituting (6.15), (6.17), and (6.18) into (6.16), we can evaluate the peak current generated by the sensor in response to the incident thermal flux:

$$i = \frac{PsdT}{dt} = \frac{Ps\Phi_s}{csh} = \frac{P}{hc}\Phi_s. \tag{6.19}$$

To establish relationship between the current and moving object, the flux from (6.14) has to be substituted into (6.19):

$$i \approx \frac{2Pa\sigma\gamma}{\pi hc} bT_a^3 \frac{\Delta T}{L^2}, \tag{6.20}$$

where $\Delta T = (T_b - T_a)$.

Several conclusions can be drawn from (6.20). The first part of the equation (the first ratio) characterizes a detector while the rest relates to an object. The pyroelectric current i is directly proportional to the temperature difference (thermal contrast) between the object and its surroundings. It is also proportional to the surface area of the object, which faces the detector. A contribution of the ambient temperature T_a is not that strong as it might appear from its third power. The ambient temperature must be entered in Kelvin, hence its variations become relatively small with respect to the scale. The thinner the sensing element the more sensitive the detector. The lens area also directly affects the signal magnitude. On the other hand, pyroelectric current does not depend on the sensor's area as long as the lens focuses the entire image on the sensing element.

To evaluate (6.20) further, let us calculate voltage across the bias resistor. That voltage can be used as indication of motions. We select a pyroelectric PVDF film sensor with typical properties: $P = 25$ μC/K m^2, $c = 2.4 \times 10^6$ J/m^3 K, $h = 25$ μm, lens area $a = 1$ cm^2, $\gamma = 0.6$, and the bias resistor $R = 10^9$ Ω (1 GΩ). We assume that the object's surface temperature is 27°C and surface area $b = 0.1$ m^2.

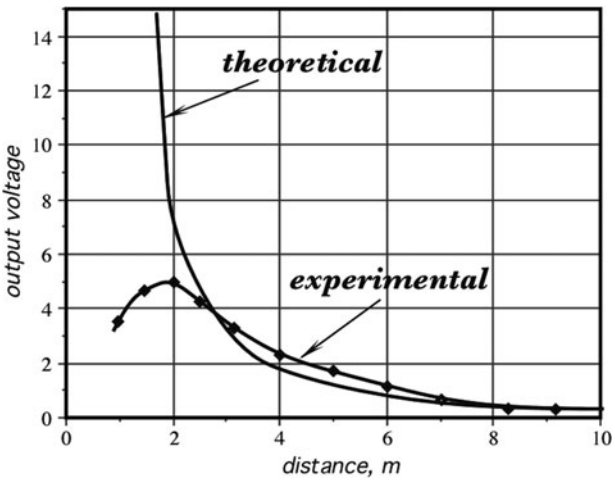

Fig. 6.19 Calculated and experimental amplitudes of output signals in a PIR detector

The ambient temperature $t_a = 20°C$. The output voltage is calculated from (6.20) as function of distance L from the detector to the object and is shown in Fig. 6.19.

The graph for Fig. 6.19 was calculated under the assumption that the optical system provides a sharp image at all distances and that image is no larger that the sensing element area. In practice, this is not always true, especially at shorter distances where the image is not only out of focus but also may overlap the out-of-phase elements of a differential sensor. A reduction in the signal amplitude at shorter distances becomes apparent; the voltage does not go as high as in the curve calculated with the above assumptions.

6.6 Optical Presence Sensors

A reflection of light is the optical phenomenon that is used quite extensively in detecting not only motion but a mere presence of an object in a monitored area. The operating principle is very simple. The sensor contains two key components: a source of light (usually a near-infrared LED) and a photodetector. The LED emits a light beam that illuminates surroundings in the field of view of the photodetector. First, the background reflection from the surrounding objects (background) is established. In terms of the output voltage from the photodetector, the light reflected from the background is V_0. A new object that appears in a foreground either absorbs more light or reflects more. In most cases, it alters the background signal by the increment ΔV, which can be detected by a threshold detector in the electronic processor. This sensor is a reflectance monitor. It will not measure a distance to the object because the value of ΔV depends on many factors, such as size

of the object, its shape, material, surface finish, and distance to the sensor. The sensor is merely a presence detector, yet in many practical cases it is just what is needed. An example of the sensor application is a presence detector for a bathroom faucet that is used to control flow of water when hands are placed under the water outlet [9, 10]. Placement of hands under the faucet controls the actuator of water flow (a solenoid-valve assembly). A similar detector is frequently employed in hand dryers, toilet tanks, light switches, robotic vacuum cleaners, and many other products.

Figures 6.20a and b show two possible locations of the sensor in a water fixture. One location is on the spout while in the other the sensor is built-in directly into the faucet. It is important to make sure that the detection area is situated where the hands are normally being placed. Figure 6.20c illustrates the faucet having a light pipe and the parts that are normally needed for dispensing water. Figure 6.21 shows a block diagram of the water flow control system. The light pipe can be a bundle of the optical fibers or a solid translucent rod molded of polycarbonate resin. Usually, the emitted light is modulated by pulses. This helps to separate the reflected light signals into a background (ambient) component and that controlled by the pulsing LED driver, since the ambient light is a d.c. or slow changing signal. Note that the background component also may contain a pulsing signal since some light will be constantly reflected from it, for example, from a sink surface. This can be taken care both the sensitivity adjustment and by detecting only a variable component of

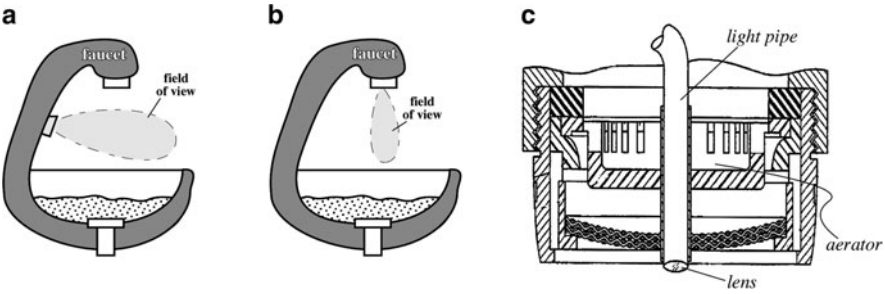

Fig. 6.20 Installation of the optical presence detector into a spout (**a**) and faucet (**b**). Cross-sectional view (**c**) of the faucet with a light pipe (adapted from Ref. [11])

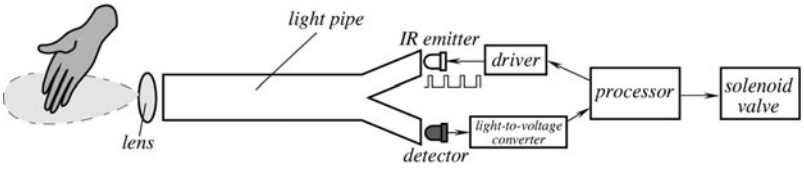

Fig. 6.21 Block diagram of the water flow controller with optical presence detector

signal ΔV. It is possible to eliminate the pilot light LED and entirely rely on the ambient illumination by independent light sources [11], however this design may be not as robust as the one with a modulated LED.

6.7 Pressure-Gradient Sensors

An efficient sensor can be designed to detect intrusion into a closed room by monitoring small variations in the atmospheric air pressure resulted from opening doors and windows or movement of people. In general, variations in air pressure can be monitored by a conventional air pressure sensor. However, this is not an efficient solution. A conventional air pressure sensor is designed for a relatively large span of the input pressures. Yet, the maximum amplitudes of the air pressure variations that are associated with intrusion are very small – over three orders of magnitude smaller than a conventional pressure sensor's span. In fact, these variations approach the noise floor of an air pressure sensor. Besides, such a sensor is just not sensitive enough for them. Appending the sensor with a high-gain amplifier is not a solution because noise will be amplified as well. The solution would be to design a sensor with a narrow pressure span but with a high sensitivity. It is also desirable to make the sensor responsive only to pressure changes, not to the absolute value of the pressure. Preferably the sensor should produce a signal similar to a first derivative of the air pressure. Since the only purpose of the sensor is to detect intrusion and not to measure the actual air pressure, accuracy requirements can be significantly relaxed for a sensor that would produce a qualitative rather than quantitative output.

A high sensitivity can be achieved by making the sensing membrane very thin with a relatively large area (see Sect. 10.4). An example of the intrusion air pressure sensor design [12] is shown in Fig. 6.22. The main part of the sensor is an enclosed chamber. The left wall of the chamber is covered with a thin stretched membrane made of a plastic or metal foil having thickness on the order of 20 μm. The membrane area should be relatively large, about 200 mm^2 or larger. The right side of the chamber is a rigid backplate with a small venting hole whose purpose is to equalize air pressures inside and outside of the chamber. The distance d from the membrane to the backplate is monitored by a built-in displacement detector (sensor). All exterior surfaces of the sensor are exposed to ambient air. When all doors and window in the monitored room are closed, the ambient air pressure is either static or changes slowly. Thanks to the vent in the backplate, pressures P_h inside and outside of the sensor's chamber are equal. When a door or window opens, the ambient air pressure changes slightly but rapidly by the value Δ. Because the vent is very narrow and air has a finite viscosity, air pressure P_h inside the chamber cannot change instantly, thus any changes inside the chamber will lag behind the outside changes. The phase lag creates a temporary air pressure differential across the membrane, which deflects to or from the backplate in relation to the differential amplitude and sign. The distance d from the membrane to the

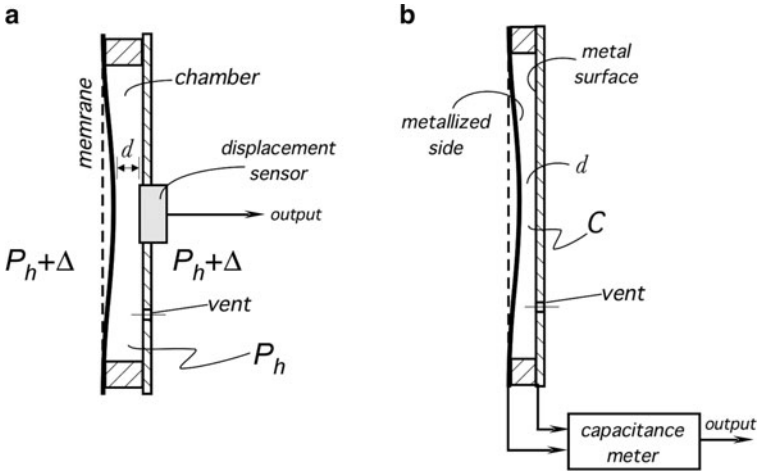

Fig. 6.22 Air pressure-gradient sensor (**a**) and the sensor with a capacitive displacement detector (**b**)

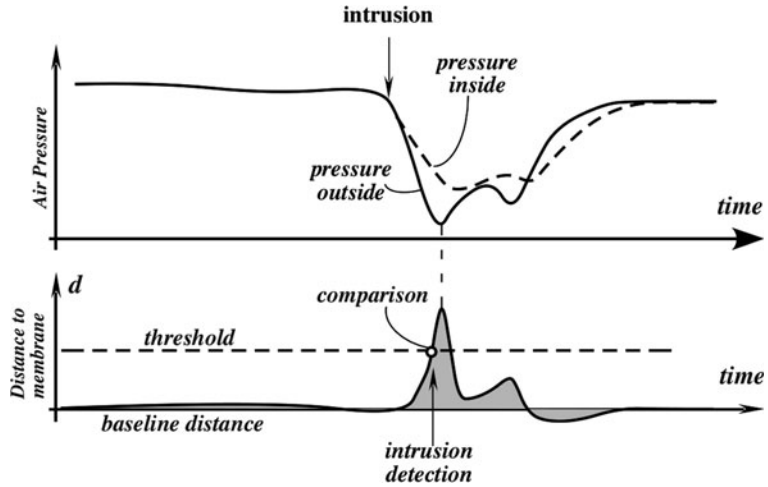

Fig. 6.23 Timing diagrams for the pressure-gradient detector

backplate is measured by a displacement sensor and is used as an indication of the intrusion. When the differential pressure is small, the membrane remains substantially flat and the distance d is at its base level. Figure 6.23 illustrates the timing diagrams of air pressures inside and outside of the sensing chamber and the differential pressure across the membrane. Note that the signal representative of the displacement d is compared with a threshold to detect an intrusion.

There are numerous ways of designing a displacement senor for monitoring the membrane deflection, many of which are discussed in Chap. 7. As an example, Fig. 6.22b illustrates a capacitive displacement sensor, where the sensing chamber

was built in form of a flat capacitor with two plates. The first plate of a capacitor is a metal foil (or metalized plastic membrane) and the other plate is a metal layer on the backplate. The baseline gap d between the membrane and backplate should be rather small: a few millimeters to the maximum. A value of the capacitance C will change when distance d varies according to the air pressure differential [see (3.20)]. The capacitance variations are measured and converted into a useful signal.

An alternative design of the above approach may include a sensor with a thermo-anemometer as a flow sensor (see Sect. 10.9) [13]. This type of a sensor can sense pressure gradients as low as few pascals or millimeters of H_2O. This is a sufficient sensitivity to detect minute air movements in a room. However, unlike the sensor of Fig. 6.22 that outputs the signal proportional to the rate of change in differential pressure to detect changes in air pressure, the sensor would require a differentiator as part of the interface circuit.

References

1. Blumenkrantz S (1989) Personal and organizational security handbook. Government Data Publications, Washington, DC
2. Ryser P, Pfister G (1991) Optical fire and security technology: sensor principles and detection intelligence. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of technical papers, pp 579–583, ©IEEE
3. Fraden J (1991) Apparatus and method for detecting movement of an object. US Patent 5,019,804, 28 May
4. Fraden J (1984) Motion discontinuance detection system and method. US Patent 4,450,351, 22 May
5. Fitz Gerald AS (1932) Photo-electric system. US Patent 2,016,032, 4 Feb 1932
6. Fraden J (1984) Toy including motion-detecting means for activating same. US Patent 4,479,329, 30 Oct
7. Fraden J (1990) Active infrared motion detector and method for detecting movement. US Patent 4,896,039, 23 Jan
8. Fraden J (1992) Active far infrared detectors. In: Temperature. Its measurement and control in science and industry, vol 6, part 2. AIP, New York, pp 831–836
9. Parsons NE et al (1999) Object-sensor-based flow-control system employing fiber-optic signal transmission. US Patent 5,984,262, 16 Nov
10. Parsons NE et al (2003) Automatic flow controller employing energy-conservation mode. US Patent 6,619,614, 16 Sept
11. Parsons NE et al (2008) Passive sensors for automatic faucets and bathroom flushers. US Patent 7,396,000, 8 July
12. Fraden J (2007) Alarm system with air pressure detector. US Patent Publication US 2008/0055079, 31 Aug
13. Fraden J (2009) Detector of low levels of gas pressure and flow. US Patent 7,490,512, 17 Feb
14. Fraden J (1988) Motion detector. US Patent 4,769,545, 6 Sept
15. Long DJ (1975) Occupancy detector apparatus for automotive safety system. US Patent 3,898,472, 5 Aug

Chapter 7
Position, Displacement, and Level

There are two ways of making progress.
One is to do something better,
the other is to do something for the first time.

The measurement of position and displacement of physical objects is essential for many applications: process feedback control, performance evaluation, transportation traffic control, robotics, security systems, just to name the few. By position, we mean determination of the object's coordinates (linear or angular) with respect to a selected reference. Displacement means moving from one position to another for a specific distance or angle. In other words, displacement is measured when an object is referenced to its own prior position rather than to an external reference.

A critical distance is measured by proximity sensors. In effect, a proximity sensor is a threshold version of a position detector. A position sensor is often a linear device whose output signal represents a distance to the object from a certain reference point. A proximity sensor, however, is a somewhat simpler device, which generates the output signal when a certain distance to the object becomes essential for an indication. For instance, many moving mechanisms in process control and robotics use a very simple but highly reliable proximity sensor, the end switch. It is an electrical switch having normally open or normally closed contacts. When a moving object activates the switch by a physical contact, the latter sends a signal to a control circuit. The signal is an indication that the object has reached the end position where the switch is positioned. Obviously, such contact switches have many drawbacks, for example, a high mechanical load on a moving object and a hysteresis.

A displacement sensor often is part of a more complex sensor where the detection of movement is one of several steps in a signal conversion. For example, Fig. 6.22 illustrates a special kind of an air pressure sensor where a variable pressure is translated into a displacement of a membrane, and the displacement is subsequently converted into an electrical signal representing pressure variations. Therefore, positions sensors, some of which are described in this chapter, are essential for designs of many other sensors which are covered in other chapters of

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_7, © Springer Science+Business Media, LLC 2010

this book. It is fair to say that position and displacement sensors are the most widely employed sensing devices.

Position and displacement sensors the zero-order devices (*see* Sect. 2.18) whose speed response usually is not critical for their performance.[1] In this section, we do not cover any sensors whose response is a function of time, which by definition, are dynamic sensors. They are covered elsewhere in this book.

When designing or selecting position and displacement detectors, the following questions should be answered:

1. How big is the displacement and of what type (linear, circular)?
2. What resolution and accuracy are required?
3. What the measured (moving) object is made of (metal, plastic, fluid, ferromagnetic, etc.)?
4. How much space is available for mounting the detector?
5. What are the environmental conditions (humidity, temperature, sources of interference, vibration, corrosive materials, etc.)?
6. How much power is available for the sensor?
7. How much mechanical wear can be expected over the life time of the machine?
8. What is the production quantity of the sensing assembly (limited number, medium volume, mass production)?
9. What is the target cost of the detecting assembly?

A careful analysis will pay big dividends in the long term.

7.1 Potentiometric Sensors

A position or displacement transducer may be built with a linear or rotary potentiometer, or a pot for short. The operating principle of this sensor is based on (3.54) for a wire resistance. From the formula, it follows that a resistance linearly relates to the wire length. Thus, by making an object to control the length of the wire, as it is done in a pot, a displacement measurement can be performed. Since a resistance measurement requires passage of electric current through the pot wire, the potentiometric transducer is an active type. That is, it requires an excitation signal, for instance, d.c. current. A moving object is mechanically coupled to the pot wiper, whose movement causes the resistance change (Fig. 7.1a). In most practical circuits, a resistance measurement is replaced by a measurement of voltage drop. The voltage across the wiper of a linear pot is proportional to the displacement d

$$v = E\frac{d}{D}, \tag{7.1}$$

[1]Nevertheless, the maximum rate of response is often specified by a manufacturer.

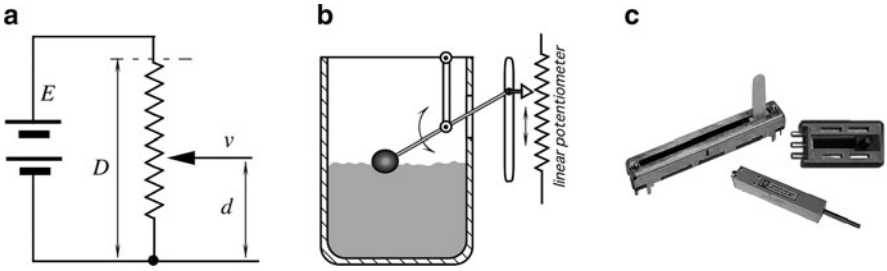

Fig. 7.1 Potentiometer as a position sensor (**a**); Fluid level sensor with a float (**b**); linear potentiometers (**c**)

where D is the full-scale displacement and E is the voltage across the pot (excitation signal). This assumes that there is no electrical loading effect from the interface circuit. If there is an appreciable load, the linear relationship between the wiper position and the output voltage will not hold. Besides, the output signal is proportional to the excitation voltage applied across the sensor. This voltage, if not maintained constant, may be a source of error. This issue may be resolved by using a ratiometric A/D converter in a microcontroller (*see* Chap. 5), so the influence of voltage will be cancelled. It should be noted that a potentiometric sensor with respect to the resistance is a ratiometric device, hence resistance of the pot is not part of the equation, as long as the resistive element is uniform over its entire length. In other words, only the ratio of resistance is important, not the resistance value. This means that the pot stability (for instance, over a temperature range) makes no effect on the accuracy. For low power applications, high impedance pots are desirable, however, the loading effect must be always considered, thus a voltage follower may be required. The wiper of the pot is usually electrically isolated from the sensing shaft. To illustrate an application of a potentiometric sensor, Fig. 7.1b shows a liquid level sensor with a float being connected to the potentiometer wiper. Different applications require different potentiometer designs, some of which are illustrated in Fig. 7.1c.

Figure 7.2a shows one problem associated with a wire-wound potentiometer. The wiper may, while moving across the winding, make contact with either one or two wires, thus resulting in uneven voltage steps (Fig. 7.2b) or a variable resolution. Therefore, when a coil potentiometer with N turns is used, only the average resolution n should be considered:

$$n = 100/N\%. \tag{7.2}$$

The force that is required to move the wiper comes from the measured object and the resulting energy is dissipated in the form of heat. Wire-wound potentiometers are fabricated with thin wires having a diameter in the order of 0.01 mm. A good coil potentiometer can provide an average resolution of about 0.1% of FS, while the high-quality resistive film potentiometers may yield an infinitesimal resolution, which is limited only by the uniformity of the resistive material and noise floor of

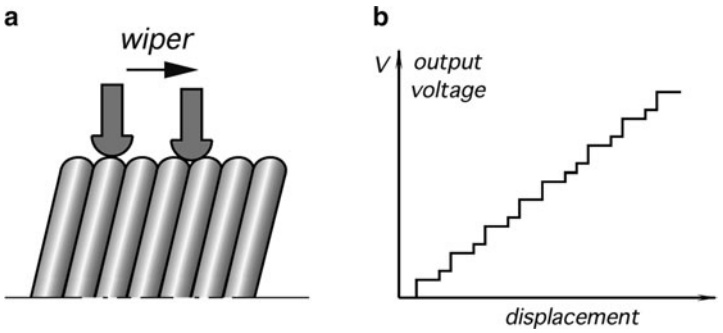

Fig. 7.2 Uncertainty caused by wire-wound potentiometer A wiper may contact one or two wires at a time (**a**); uneven voltage steps (**b**)

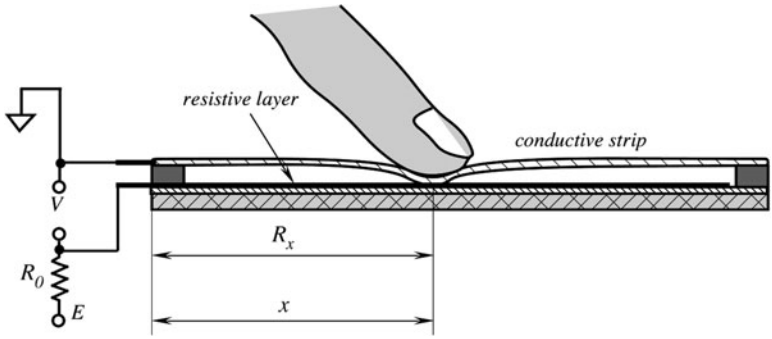

Fig. 7.3 Principle of a pressure-sensitive potentiometric position sensor

the interface circuit. The continuous resolution pots are fabricated with conductive plastic, carbon film, metal film, or a ceramic-metal mix, which is known as *cermet*. The wiper of the precision potentiometers is made from precious metal alloys. Displacements sensed by the angular potentiometers range from approximately 10° to over 3,000° for the multiturn pots (with gear mechanisms).

A concept of another implementation of a potentiometric position sensor with a continuous resolution is shown in Fig. 7.3. The sensor consists of two strips – the upper strip is made of flexible plastic sheet having a metalized surface. This is a contact or wiper strip. The bottom strip is rigid and coated with a resistive material of a total resistance ranging from several kΩ to megohms. The upper conductive strip (wiper) and the bottom resistive strip are connected into an electric circuit. When a pusher (e.g. a finger) is pressed against the upper strip at a specific distance x from the end, the contact strip touches the bottom strip and makes en electric contact at the pressure point. That is, the contact strip works as a wiper in a pot. The contact between two strips changes the output voltage from E to ER_x/R_0, which is proportional to distance x from the left side of the sensor. The pusher (wiper) may slide along the sensor causing a variable output voltage.

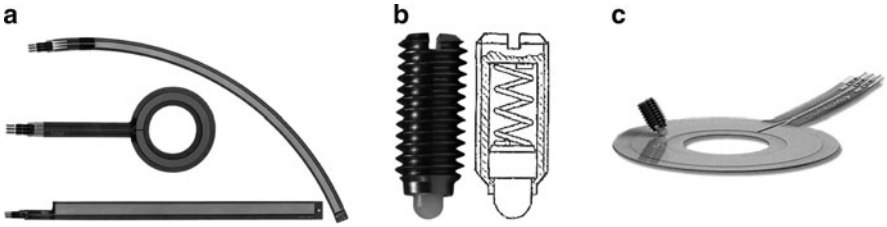

Fig. 7.4 Various shapes of pressure sensitive potentiometers (**a**), a wiper with a spring-loaded tip (**b**) and a circular SoftPot with a wiper (**c**) (Courtesy of spectrasymbol.com)

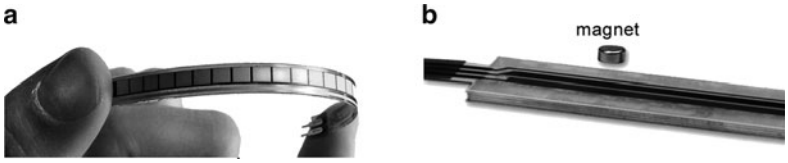

Fig. 7.5 Flex sensor (**a**) and MagnetoPot (**b**) (courtesy of spectrasymbol.com)

Practical examples of various shapes of the pressure-sensitive potentiometers are shown in Fig. 7.4 where a resistive layer is deposited on a polyester substrate. The overall resistance varies from 1 to 100 kΩ with a wiper (pusher) force is in the range from 1 to 3 N. Note that the wipers should be fabricated on a slippery material, such as derlin or nylon. Alternatively, a roller can be used as a wiper.

Another interesting potentiometric sensor uses the piezoresistive properties of carbon-impregnated plastics. The operating principle of such a sensor is based on change in resistance in response to a mechanical deformation. Carbon-impregnated layer is deposited on a substrate that is fabricated of polyester, fiberglass or polyimide. When deformed, the carbon particles density varies and subsequently varies the overall resistance. It is the same principle that used in strain gauges and is the basis of a flex sensor (Fig. 7.5a). Such sensors may be used for motion control, medical devices, musical instruments, robotics, and other devices where bends or mutual rotation of parts have to be monitored. There are two main brands of the flex sensors that are affordable and easily available: SpectraSymbol Flex sensors and the Gentile-Abrams sensor, available from Jameco. It should be noted that the bend sensors change resistance to any deformation of the substrate, including local multiple bends and linear stresses. Also, they possess a very noticeable hysteresis and may be the source of noise.

While being quite useful in many applications, potentiometers with contact wipers have several drawbacks:

1. Noticeable mechanical load (friction)
2. Need for a physical coupling with the object
3. Low speed
4. Friction and excitation voltage cause heating of the potentiometer
5. Low environmental stability (wear, susceptibility to dust)

A potentiometric sensor where the physical contact between the wiper and a resistive layer and thus friction are eliminated uses a wiper layer impregnated with ferromagnetic particles. When an external magnetic filed is present at a specific location above the potentiometer, the contact layer is pulled up to the conductive layer making an electrical contact, just like the wiper (Fig. 7.5b). This magnetic potentiometer is sealed so it can be used as an immersed sensor for measuring, for example, level of a liquid. It is important to select the appropriate magnet that is sufficiently strong for the sensor operation (*see* Sect. 3.3.4). Even though such a contactless potentiometer has no friction, a magnetic drag is still a force that opposes the magnet motion. That force should be considered and accounted for in sensitive applications.

7.2 Capacitive Sensors

The capacitive displacement sensors have very broad applications – they are employed directly to gauge displacement and position and also as building blocks in other sensors where displacements is produced by force, pressure, temperature, etc. The ability of capacitive detectors to sense virtually all materials makes them an attractive choice for many applications. Equation (3.20) defines that the capacitance of a flat capacitor is inversely proportional to the distance between the plates. The operating principle of a capacitive gauge , proximity , and position sensors is based on either changing the geometry (i.e. a distance between the capacitor plates), or capacitance variations in the presence of conductive or dielectric materials. When the capacitance changes, it can be converted into a variable electrical signal. As with many sensors, a capacitive sensor can be either monopolar (using just one capacitor), differential (using two capacitors), or a capacitive bridge can be employed (using four capacitors). When two or four capacitors are used, one or two capacitors may be either fixed or variable with the opposite phase.

As an introductory example, consider three equally spaced plates where each has area A (Fig. 7.6a). The plates form two capacitors C_1 and C_2. The upper and lower plates are fed with the out-of-phase sinewave signals, that is, the signal phases are shifted by 180°. Both capacitors nearly equal one another and thus the central plate has almost no voltage with respect to ground – the charges on C_1 and C_2 cancel each

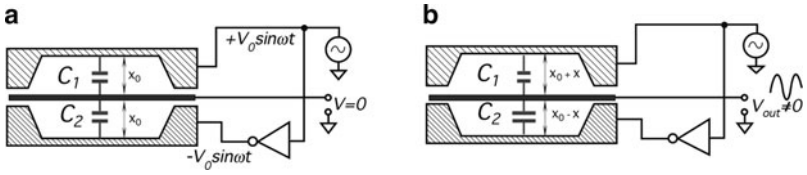

Fig. 7.6 Operating principle of a flat-plate capacitive sensor. Balanced position (**a**) and disbalanced position (**b**)

other. Now, let us assume that the central plate moves downward by distance x (Fig. 7.6b). This causes changes in the respective capacitance values:

$$C_1 = \frac{\varepsilon A}{x_0 + x} \quad \text{and} \quad C_2 = \frac{\varepsilon A}{x_0 - x}, \tag{7.3}$$

and the central plate signal increases in proportion to the displacement while the phase of that signal is the indication of the central plate direction – up or down. The amplitude of the output signals is

$$V_{out} = V_0 \left(-\frac{x}{x_0 + x} + \frac{\Delta C_0}{C_0} \right) \tag{7.4}$$

As long as $x << x_0$, the output voltage may be considered a linear function of displacement. The second summand represents an initial capacitance mismatch and is the prime cause for the output offset. The offset is also caused by the fringing effects at the peripheral portions of the plates and by the so-called electrostatic force. The force is a result of the charge attraction and repulsion which is applied to the plates of the sensor and the plates behave like springs. The instantaneous value of the force is

$$F = -\frac{1}{2} \frac{CV^2}{x_0 + x}. \tag{7.5}$$

In another design, two separate plates are fabricated by using a MEMS technology (Fig. 7.7). The plates are macromachined of silicon [21]. One plate serves for a displacement measurement, while the other is for reference. Both plates have nearly the same surface areas, however the measurement plate is supported by four flexible suspensions, while the reference plate is held by the stiff suspensions. This particular design is especially useful for accelerometers.

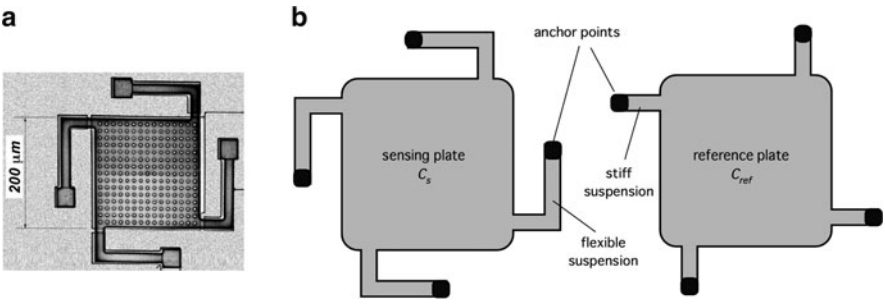

Fig. 7.7 A dual-plate capacitive displacement sensor (adapted from [22]). Micromachined sensing plate (**a**) and different suspensions for the sensing and reference plates (**b**)

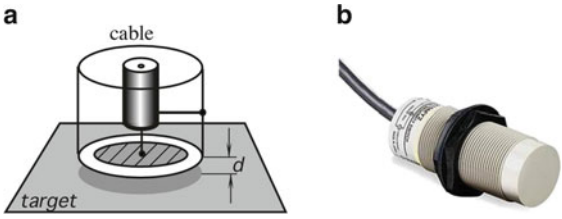

Fig. 7.8 Capacitive probe with a guard ring cross-sectional view (**a**); outside view (**b**)

In many practical applications, when measuring distances to an electrically conductive object, the object's surface itself may serve as a capacitor's plate. A design of a monopolar capacitive sensor is shown in Fig. 7.8, where one plate of a capacitor is connected to the central conductor of a coaxial cable, while the other plate is formed by a target (object). Note that the probe plate is surrounded by a grounded guard to minimize a fringing effect and improve linearity. A typical capacitive probe operates at frequencies in the 3 MHz range and can detect very fast moving targets, since a frequency response of a probe with a built-in electronic interface is in the range of 40 kHz. A capacitive proximity sensor can be highly efficient when used with electrically conductive objects. The sensor measures a capacitance between the electrode and the object. Nevertheless, even for the nonconductive objects, these sensors can be employed quite efficiently though with lesser accuracy. Any object, conductive or nonconductive, that is brought in the vicinity of the electrode, has its own dielectric properties that will alter the capacitance between the electrode and the sensor housing and, in turn, will produce a measurable response.

To improve sensitivity and reduce fringing effects, the monopolar capacitive sensor may be supplied with a driven shield. The idea behind a driven shield is to eliminate the electric field between the sensing electrode and undesirable parts of the object, thus making the parasitic capacitance being virtually nonexistent. A driven shield is positioned around the nonoperating sides of the electrode and fed with voltage equal to that at the electrode. Since the shield and the electrode voltages are in-phase and have the same magnitude, no electric field exists between the two and all components positioned behind the shield make no effect on the operation. The driven shield technique is illustrated in Fig. 7.9.

Nowadays, a capacitive bridge becomes increasingly popular in designs of the displacement sensors [1]. A linear bridge capacitive position sensor [3] is shown in Fig. 7.10a. The sensor comprises two planar electrode sets that are parallel and adjacent to each other with a constant separation distance, d. The increase in capacitance, and the spacing between the plate sets is relatively small. The stationary electrode set contains four rectangular elements while the moving electrode set contains two rectangular elements. All six elements are of about the same size (a side dimension is b). The size of each plate can be as large as is mechanically practical, when a large range of linearity is desired. The four electrodes of the stationary set are cross-connected electrically thus forming a bridge type capacitance network.

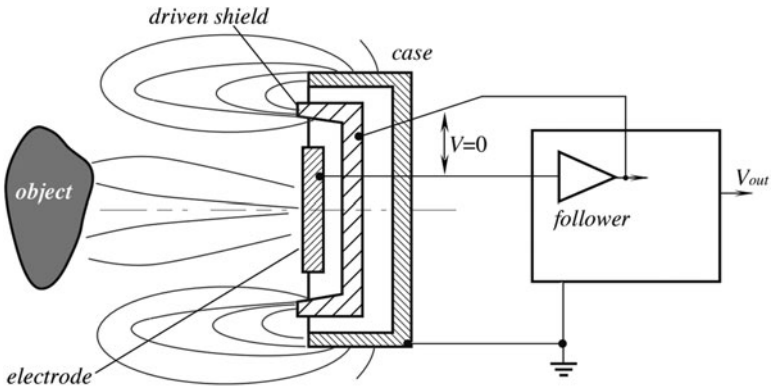

Fig. 7.9 Driven shield around the electrode in a capacitive proximity sensor

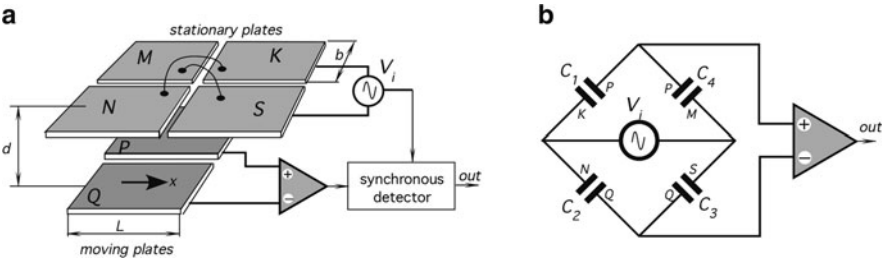

Fig. 7.10 Parallel-plate capacitive bridge sensor plate arrangement (**a**) and equivalent circuit diagram (**b**)

A bridge excitation source provides a sinusoidal voltage (5–50 kHz) and the voltage difference between the pair of moving plates is sensed by the differential amplifier whose output is connected to the input of a synchronous detector. The capacitance of two parallel plates at fixed separation distance is proportional to the area of either plate that directly faces the corresponding area of the other plate. Figure 7.10b shows the equivalent circuit of the sensor that has a configuration of a capacitive bridge. A value of capacitor C_1 is

$$C_1 = \frac{\varepsilon_0 b}{d}\left(\frac{L}{2} + x\right) \tag{7.6}$$

The other capacitances are derived from the identical equations. Note that the opposite capacitors are nearly equal: $C_1 = C_3$ and $C_2 = C_4$. A mutual shift of the plates with respect to a fully symmetrical position results in the bridge disbalance and the phase-sensitive output of the differential amplifier. An advantage of the capacitive bridge circuit is the same as of any bridge circuit: linearity and external noise immunity. Besides the flat electrodes as described above, the same method

can be applied to any symmetrical arrangement of the sensor, for instance, to detect a rotary motion.

7.3 Inductive and Magnetic Sensors

One of many advantages of using magnetic field for sensing position and distance is that any nonmagnetic material can be penetrated by the filed with no loss of position accuracy. Stainless steel, aluminum, brass, copper, plastics, masonry, and woods can be penetrated, meaning that accurate position with respect to the probe at the opposite side of a wall can be determined nearly instantly. Another advantage is that the magnetic sensors can work in severe environments and corrosive situations because the probes and targets can be coated with inert materials that will not adversely affect the magnetic fields.

7.3.1 LVDT and RVDT

Position and displacement may be sensed by methods of electromagnetic induction. A magnetic flux coupling between two coils may be altered by the movement of an object and subsequently converted into voltage. Variable inductance sensors that use a nonmagnetized ferromagnetic medium to alter the reluctance (magnetic resistance) of the flux path are known as variable-reluctance transducers [3]. The basic arrangement of a multi-induction transducer contains two coils: primary and secondary. The primary carries ac excitation (V_{ref}) that induces a steady a.c. voltage in the secondary coil (Fig. 7.11). The induced amplitude depends on the flux coupling between the coils. There are two techniques to change the coupling. One is the movement of an object made of ferromagnetic material within the flux path. This changes the reluctance of the path, which, in turn, alters the coupling between the coils. This is the basis for the operation of linear variable differential transformer (LVDT), rotary variable differential transformer (RVDT), and the mutual inductance proximity sensors. The other method is to physically move one coil with respect to another.

LVDT is a transformer with a mechanically actuated core. The primary coil is driven by a sine wave (excitation signal) having a stabilized amplitude. Sine wave eliminates error related harmonics in the transformer [4]. An ac signal is induced in the secondary coils. A core made of a ferromagnetic material is inserted coaxially into the cylindrical opening without physically touching the coils. The two secondaries are connected in the opposed phase. When the core is positioned in the magnetic center of the transformer, the secondary output signals cancel and there is no output voltage. Moving the core away from the central position unbalances the induced magnetic flux ratio between the secondaries, developing an output. As the core moves, the reluctance of the flux path changes. Hence, the degree of flux

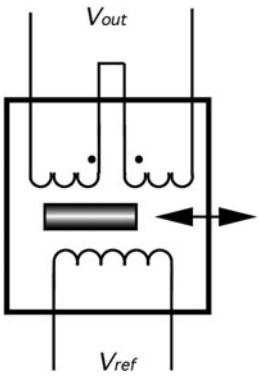

Fig. 7.11 Circuit diagram of the LVDT sensor

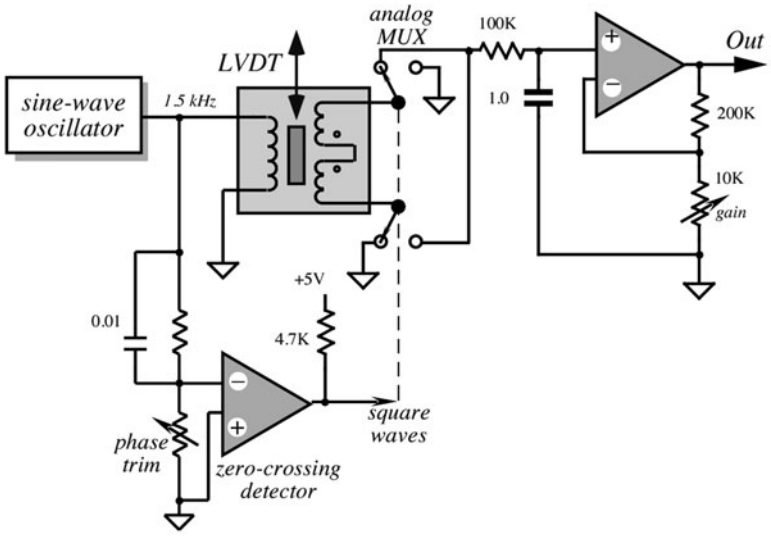

Fig. 7.12 A simplified circuit diagram of an interface for an LVDT sensor

coupling depends on the axial position of the core. At a steady state, the amplitude of the induced voltage is proportional, in the linear operating region, to the core displacement. Consequently, voltage may be used as measure of a displacement. The LVDT provides the direction as well as magnitude of the displacement. The direction is determined by the phase angle between the primary (reference) voltage and the secondary voltage. Excitation voltage is generated by a stable oscillator. To exemplify how the sensor works, Fig. 7.12 shows the LVDT connected to a synchronous detector that rectifies the sine wave and presents it at the output as a d.c. signal. The synchronous detector is comprised of an analog multiplexer (MUX) and a zero-crossing detector, which converts the sine wave into the square

pulses compatible with the control input of the multiplexer. A phase of the zero-crossing detector should be trimmed for the zero output at the central position of the core. The output amplifier can be trimmed to a desirable gain to make the signal compatible with the next stages. The synchronized clock to the multiplexer means that the information presented to the *RC*-filter at the input of the amplifier is amplitude- and phase-sensitive. The output voltage represents how far the core is from the center and on which side.

For LVDT to measure transient motions accurately, the frequency of the oscillator must be at least ten times higher than the highest significant frequency of the movement. For a slow-changing process, stable oscillator may be replaced by coupling to a power line frequency of 60 or 50 Hz.

The advantages of the LVDT and RVDT are the following: (1) the sensor is a noncontact device with no or very little friction resistance with small resistive forces; (2) hysteresis (magnetic and mechanical) are negligible; (3) output impedance is very low; (4) low susceptibility to noise and interferences; (5) construction is solid and robust; and (6) infinitesimal resolution is possible.

One useful application for the LVDT sensor is in the so-called gauge heads, which are used in tool inspection and gauging equipment. In that case, the inner core of the LVDT is spring loaded to return the measuring head to a preset reference position.

The RVDT operates on the same principle as LVDT, except that a rotary ferromagnetic core is used. The prime use for the RVDT is the measurement of angular displacement. A typical linear range of measurement is about $\pm$ 40° with a nonlinearity error of about 1%.

7.3.2 Eddy Current Sensors

Eddy current is an electrical phenomenon discovered in 1851 by French physicist Léon Foucault. Thus, this current sometimes is called *Foucault current*. It appears in two cases: (1) when a conductor is exposed to a changing magnetic field due to relative motion of the field source and conductor and (2) due to changing intensity of the magnetic field. These effects cause a circulating flow of electrons, or a current, within the body of the conductor. These circulating eddies of current create induced magnetic fields that oppose the change of the original magnetic field due to Lenz's law, causing repulsive or drag forces between the conductor and the magnet. The stronger the applied magnetic field, or the greater the electrical conductivity of the conductor, or the faster the field that the conductor is exposed to changes, then the greater the currents that are developed and the greater the opposing field.

To sense proximity of nonmagnetic but conductive materials, the effect of eddy currents is employed in a dual-coil sensor (Fig. 7.13a). One coil is used as a reference, while the other is for sensing the magnetic currents induced in the conductive object. Eddy currents produce the magnetic field, which opposes that of the sensing coil, thus resulting in a disbalance with respect to the reference coil.

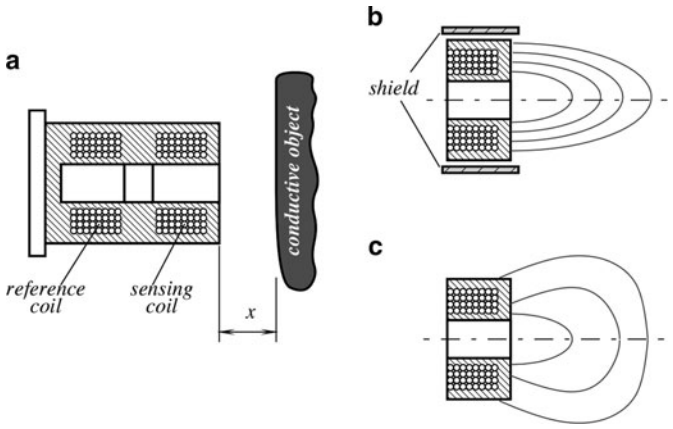

Fig. 7.13 Electromagnetic proximity sensor with eddy currents (**a**). Sensor with the shielded front end (**b**); Unshielded sensor (**c**)

The closer the object to the coil the larger the change in the magnetic impedance. The depth of the object where eddy currents are produced is defined by

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}, \tag{7.7}$$

where f is the frequency and σ is the target conductivity. Naturally, for the effective operation, the object thickness should be larger than that depth. Hence, eddy detectors should not be used for detecting film metallized or foil objects. Generally, the relationship between the coil impedance and distance to the object x is nonlinear and temperature-dependent. The operating frequency range of the eddy current sensors ranges from 50 kHz to 10 MHz.

Figure 7.13b and c show two configurations of the eddy current sensors: with and without the shield. The shielded sensor has a metal guard around the ferrite core and the coil assembly. It focuses and directs the electromagnetic field to the front of the sensor. This allows the sensor to be installed and even imbedded into a metal structure without influencing the detection range. The unshielded sensor can sense at its sides as well as from the front. As a result, the detecting range of an unshielded sensor is usually somewhat greater than that of the shielded sensor of the same diameter, however, to operate properly, the unshielded sensors require nonmetallic surrounding objects.

In addition to position detection, eddy sensors can be used to determine material thickness, nonconductive coating thickness, conductivity and plating measurements, and cracks in the material. Crack detection and surface flaws become the most popular applications for the sensors. Cracks interrupt flow of eddy currents and result in abrupt change in the sensor's output signal.

Depending on the applications, eddy probes may be of many coil configurations: very small in diameter (2–3 mm) or quite large (25 mm). Some companies even make custom-designed probes to meet unique requirements of the customers

(www.olympus-ims.com). One important advantage of the eddy current sensors is that they do not need magnetic material for the operation, thus they can be quite effective at high temperatures (well exceeding Curie temperature of a magnetic material), and for measuring a distance to or level of conductive liquids, including molten metals. Another advantage of the detectors is that they are not mechanically coupled to the object and thus the loading effect is very low.

7.3.3 Transverse Inductive Sensor

Another electromagnetic position-sensing device is called a transverse inductive proximity sensor. It is useful for sensing relatively small displacements of ferromagnetic materials. As the name implies, the sensor measures the distance to an object, which alters the magnetic field in the coil. The coil inductance is measured by an external electronic circuit (Fig. 7.14). A self-induction principle is the foundation for the operation of the sensor. When it moves into the vicinity of a ferromagnetic object, its magnetic field changes, thus altering the inductance of the coil. The advantage of the sensor is that it is a noncontact device whose interaction with the object is only through the magnetic field. An obvious limitation is that it is useful only for the ferromagnetic objects at relatively short distances.

A modified version of the transverse transducer is shown in Fig. 7.15a. To overcome the limitation for measuring only ferrous materials, a ferromagnetic disk is attached to a displacing object while the coil has a stationary position. Alternatively, the coil may be attached to the object and the core is stationary. This proximity sensor is useful for measuring small displacements only, as its linearity is poor in comparison with LVDT. However, it is quite useful as a proximity detector for the indication of close proximity to an object, which is made of any solid material. A magnitude of the output signal as function of distance to the disk is shown in Fig. 7.15b.

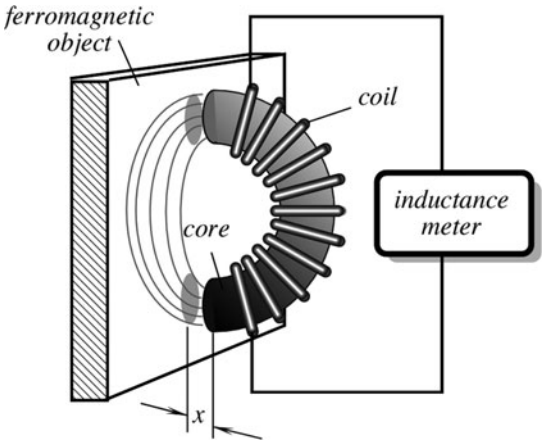

Fig. 7.14 Transverse inductive proximity sensor

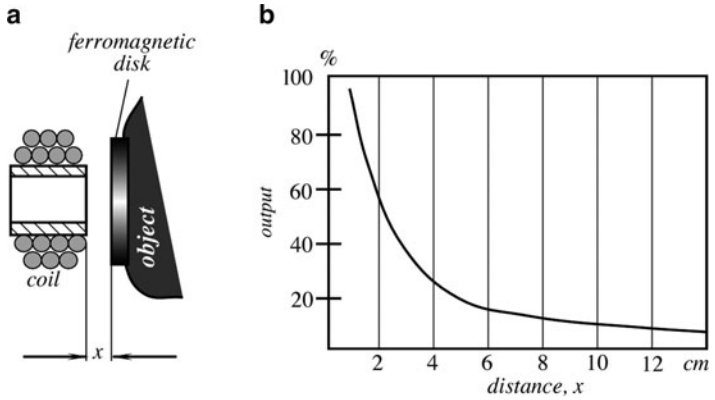

Fig. 7.15 Transverse sensor with an auxiliary ferromagnetic disc (**a**) and the output signal as function of distance (**b**)

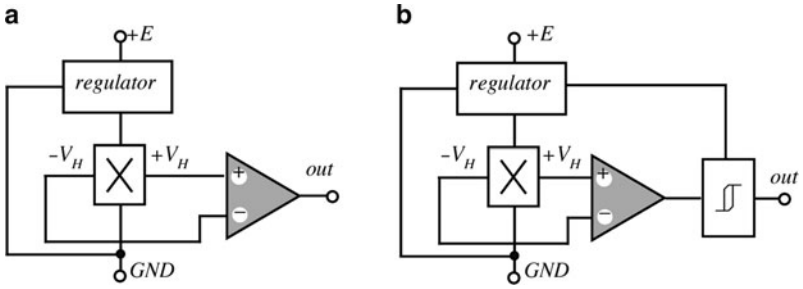

Fig. 7.16 Circuit diagrams of a linear (**a**) and threshold (**b**) Hall effect sensors

7.3.4 Hall Effect Sensors

Probably the most widely used magnetic sensors are the Hall effect sensors.[2] There are two types of the Hall sensors: analog and bi-level (Fig. 7.16). Analog sensors usually incorporate amplifiers for easier interface with the peripheral circuits. In comparison with a basic sensor (Fig. 3.30), the analog sensor operates over a broader voltage range and more stable in a noisy environment. These sensors are not quite linear (Fig. 7.17a) with respect to the magnetic field density and, therefore, for precision measurements require a calibration. A bi-level sensor in addition to the amplifier contains a Schmitt trigger with a built-in hysteresis of the threshold level. The output signal as function of a magnetic field density is shown in Fig. 7.17b. The signal is two-level and has clearly pronounced hysteresis with respect to the magnetic field. When the applied magnetic flux density exceeds a certain threshold, the trigger provides a clean transient from the OFF to ON

[2]See Sect. 3.8 for the operating principle.

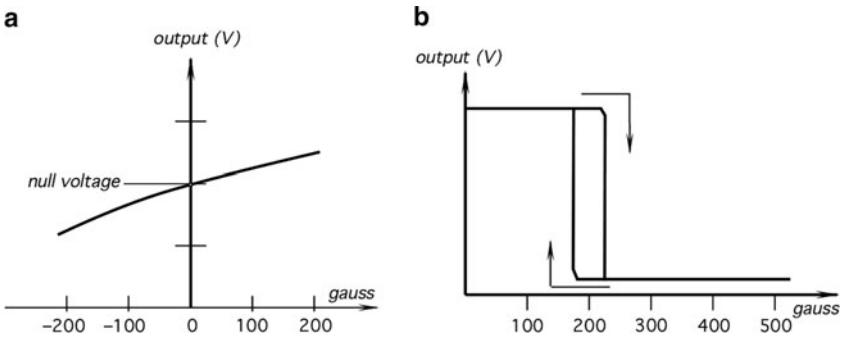

Fig. 7.17 Transfer functions of a linear (**a**) and a threshold (**b**) Hall effect sensors

position. The hysteresis eliminates spurious oscillations by introducing a dead band zone in which the action is disabled after the threshold value has passed. The Hall sensors are usually fabricated as monolithic silicon chips and encapsulated into small epoxy or ceramic packages.

For position and displacement measurements, the Hall effect sensors must be provided with a magnetic field source and interface electronic circuit. A magnetic field has two important for this application characteristics: flux density and polarity (or orientation). It should be noted that for better responsivity, magnetic field lines must be normal (perpendicular) to the flat face of the sensor and must be at a correct polarity. For example, in the bi-level sensors fabricated by Sprague®, the south magnetic pole will cause switching action and the north pole will have no effect.

Before designing a position detector with a Hall sensor, an overall analysis should be performed in approximately the following manner. First, the field strength of the magnet should be investigated. The strength will be the greatest at the pole face, and will decrease with increasing distance from the magnet. The field may be measured by the gaussmeter or a calibrated analog Hall sensor. For a bi-level type Hall sensor, the longest distance at which the sensor's output goes from ON (high) to OFF (low) is called a release point. It can be used to determine a critical distance where the sensor is useful. A magnetic field strength is not linear with distance and depends greatly upon magnet shape, magnetic circuit, and path traveled by the magnet. The Hall conductive strip is situated at some depth within the sensor's housing. This determines the minimum operating distance. A magnet must operate reliably with the total effective air gap in the working environment. It must fit the available space, must be mountable, affordable, and available.[3]

As a first example of the Hall sensor application, let us look at a liquid level detector with a float (Fig. 7.18). A permanent magnet is imbedded inside a float having a hole in the center. The float can freely slide up and down over the pole that is positioned inside the tank containing liquid. The float position corresponds to the

[3]For more information on permanent magnets see Sect. 3.3.4.

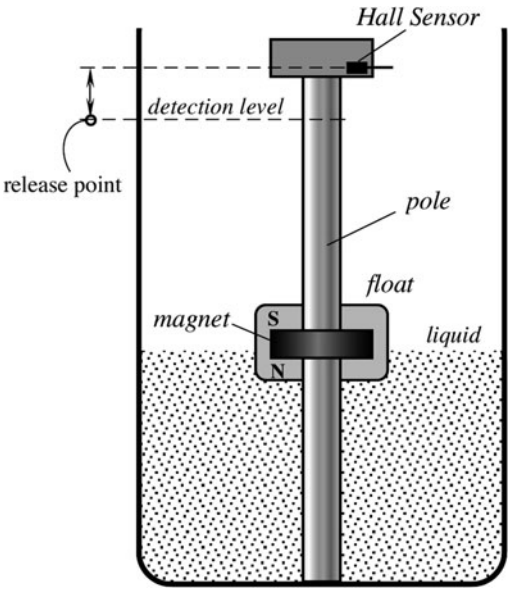

Fig. 7.18 Liquid level detector with a Hall sensor

liquid surface level. A bi-level Hall sensor is mounted at the top of the pole, which should be fabricated on a nonmagnetic material. When the liquid level rises and reaches the detection level (release point), the Hall switch triggers and sends signal to the monitoring device. When the liquid level drops below the release point plus the threshold hysteresis, the Hall sensor output voltage changes, indicating that the liquid level dropped. The detection point depends on three key factors: the magnet strength and shape, the Hall sensor's sensitivity, the hysteresis, and presence of ferromagnetic components in the vicinity of the Hall sensor.

The Hall sensors can be used for interrupter switching with a moving object. In this mode, the activating magnet and the Hall sensor are mounted on a single rugged assembly with a small air gap between them (Fig. 7.19). Thus, the sensor is held in the ON position by the activating magnet. If a ferromagnetic plate, or vane, is placed between the magnet and the Hall sensor, the vane forms a magnetic shunt that distorts the magnetic flux away from the sensor. This causes the sensor to flip to the OFF position. The Hall sensor and the magnet could be molded into a common housing, thus eliminating the alignment problem. The ferrous vanes, which interrupt the magnetic flux, could have linear or rotating motion. An example of such a device is an automobile distributor.

Like many other sensors, four Hall sensors can be connected into a bridge circuit to detect linear or circular motion. Figures 7.20a and b illustrate this concept where the sensor is fabricated using MEMS technology on a single chip and packaged in a SOIC-8 plastic housing. A circular magnet is positioned above the chip and its angle of rotation and direction is sensed and converted into a digital code. The

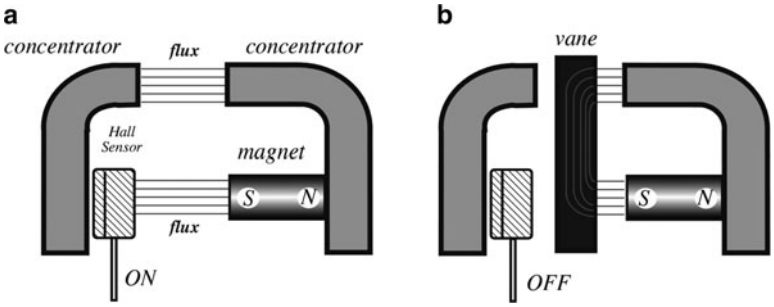

Fig. 7.19 The Hall effect sensor in the interrupter switching mode
The magnetic flux turns the sensor on (**a**); the magnetic flux is shunted by a vane (**b**) (after [5])

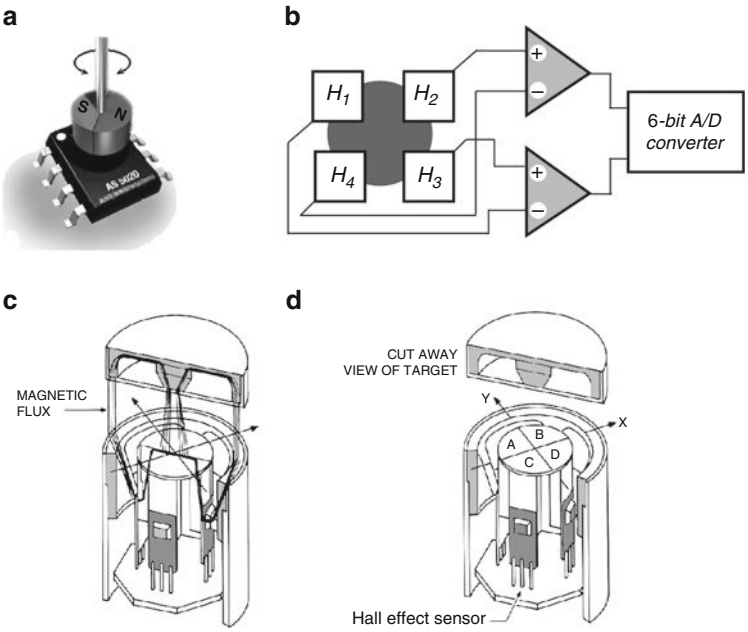

Fig. 7.20 Angular Hall-sensor bridge (**a**) and the internal sensor interface (**b**) (courtesy of Austria Micro Systems).
A cutaway view (**c**) of the sensor with the target and the probe shows the magnetic flux paths. A cut-away view (**d**) shows four Hall effect sensors with four flux return path

properties of an analog-to-digital converter determine the speed response that allows the magnet to rotate with a rate of up to 30,000 rpm. Such a sensor permits a friction-free precision linear and angular sensing of position, precision angular encoding, and even to make a programmable rotary switch. Thanks to a bridge connection of the individual sensor, the circuit is highly tolerant to the magnet misalignment and external interferences, including the magnetic fields.

The design of a 3-D coordinate Hall effect sensor works by electronically measuring and comparing the magnetic flux from a movable target through four geometrically equal magnetic paths arranged symmetrically around the axis of the probe (Fig. 7.20c, d). It is a magnetic equivalent of a Wheatstone bridge. The target's symmetrical magnetic field, generated by a permanent magnet, travels from the central pole through the air to the outer rim when it is not in the vicinity of the probe. Because the flux from the target will take the path of least resistance (reluctance), the flux will go through the probe when the target is sufficiently close to it. The probe has a central pole face divided into four equal sections. The values of flux in the *A, B, C*, and *D* paths are measured by the respective Hall effect sensors. There are two ways to fabricate a target. One is active and the other is passive. An active target uses a permanent magnet to generate a magnetic field, which is sensed by the probe when it is within the operating range. A passive target does not generate magnetic field, instead, the field is generated by the probe and returned by the target. An example of the application is the unmanned vehicle guidance system that leads a vehicle over a roadbed with a passive metal strip targets buried just under the road surface. The probe is attached to the vehicle. The targets will give position, speed, and direction as the probe passes over it. A probe and target can be separated by several inches.

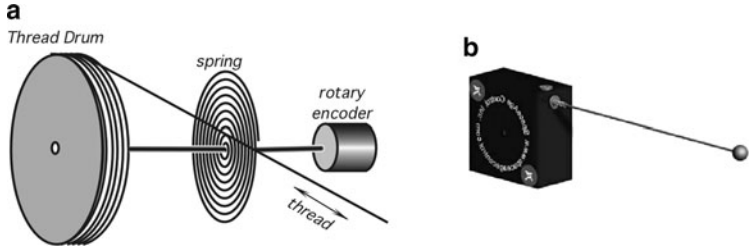

Fig. 7.21 Conversion of a linear displacement (length of a thread or cable) into a rotary motion (**a**) and cable position sensor (**b**) (courtesy of SpaceAge Control, Inc.)

A rotary motion can be digitally encoded with high precision. To take advantage of this feature, a linear distance sensor can be built with a converter of a linear into a rotary motion as shown in Fig. 7.21. Such sensors are produced, for example, by SpaceAge Control, Inc. (www.spaceagecontrol.com).

7.3.5 Magnetoresistive Sensors[4]

These sensors are similar in application to the Hall effect sensors. For functioning, they require an external magnetic field. Hence, whenever the magnetoresistive sensor is used as a proximity, position, or rotation detector, it must be combined with a source

[4]Information on the KZM10 and KM110 sensors is courtesy of Philips Semiconductors BV. Eindhoven, The Netherlands.

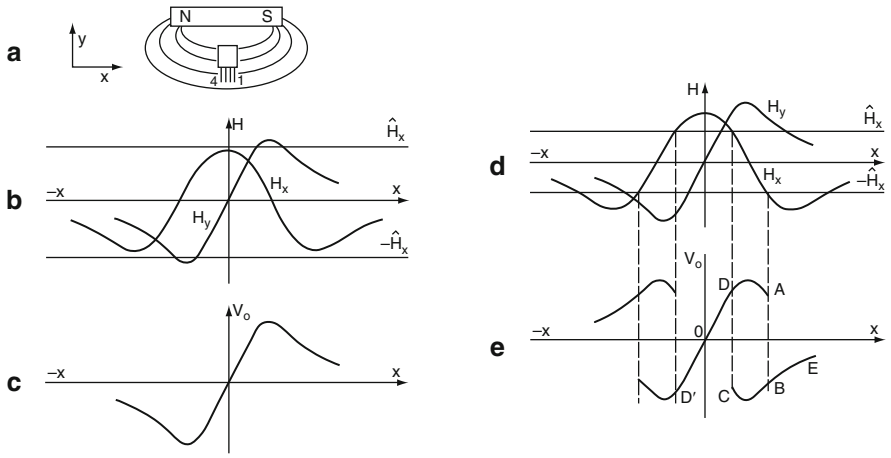

Fig. 7.22 Magnetoresistive sensor output in the field of a permanent magnet as a function of its displacement x parallel to the magnetic axis (**a**–**c**). The magnet provides both the axillary and transverse fields. Reversal of the sensor relative to the magnet will reverse the characteristic. Sensor output with a too strong magnetic field (**d** and **e**)

of a magnetic field. Usually, the field is originated in a permanent magnet, which is attached to the sensor. Figure 7.22 shows a simple arrangement for using a sensor-permanent-magnet combination to measure linear displacement. It reveals some of the problems likely to be encountered if proper account is not taken of the effects described below. When the sensor is placed in the magnetic field, it is exposed to the fields in both the x and y directions. If the magnet is oriented with its axis parallel to the sensor strips (i.e. in the x-direction) as shown in Fig. 7.22a, $\mathbf{H}_x$ then provides the auxiliary field, and the variation in $\mathbf{H}_y$ can be used as a measure of x displacement. Figure 7.22b shows how both $\mathbf{H}_x$ and $\mathbf{H}_y$ vary with x, and Fig. 7.22c shows the corresponding output signal. In this example, $\mathbf{H}_x$ never exceeds $\pm\hat{\mathbf{H}}_x$ (the field that can cause flipping of the sensor) and the sensor characteristics remain stable and well-behaved throughout the measuring range. However, if the magnet it too powerful, or the sensor passes too close to the magnet, the output signal will be drastically different.

Suppose the sensor is initially on the transverse axis of the magnet ($x = 0$). $\mathbf{H}_y$ will be zero and $\mathbf{H}_x$ will be at its maximum value ($>\mathbf{H}_x$). So the sensor will be oriented in the +x direction and the output voltage will vary as in Fig. 7.21e. With sensor's movement in +x direction, $\mathbf{H}_y$ and V_0 increase, and $\mathbf{H}_x$ falls to zero and then increases negatively until $\mathbf{H}_y$ exceeds $-\mathbf{H}_x$. At this point, the sensor characteristic flips and the output voltage reverses, moving from A to B in Fig. 7.22e. A further increase in x causes the sensor voltage to move along BE. If the sensor is moved in the opposite direction, however, $\mathbf{H}_x$ increases until it exceeds $+\mathbf{H}_x$ and V_0 moves from B to C. At this point, the sensor characteristic again flips and V_0 moves from C to D. Then, under these conditions, the sensor characteristic will trace the hysteresis loop ABCD, and a similar loop in the $-$x direction. Figure 7.22e is an idealized case, since the reversals are never as abrupt as shown.

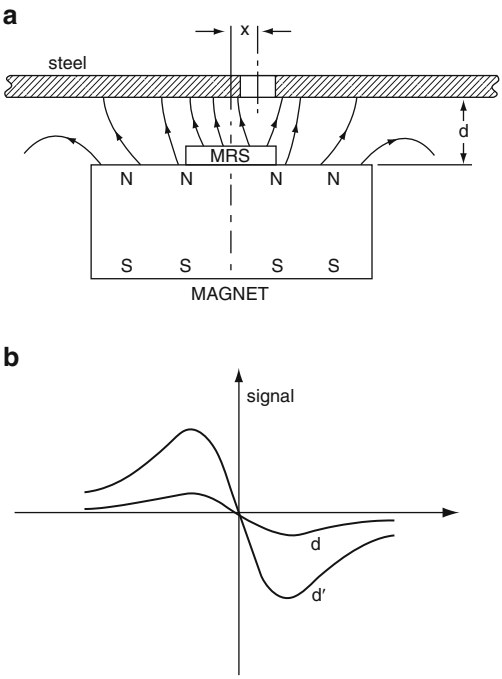

Fig. 7.23 One point measurement with the KMZ10. The sensor is located between the permanent magnet and the metal plate (**a**). Output signals for two distances between the magnet and the plate (**b**)

Figure 7.23a shows how KMZ10B and KM110B magnetoresistive sensors may be used to make position measurements of a metal object. The sensor is located between the plate and a permanent magnet, which is oriented with its magnetic axis normal to the axis of the metal plate. A discontinuity in the plate's structure, such as a hole or a region of nonmagnetic material, will disturb the magnetic field and produce a variation in the output signal from the sensor. Figure 7.23b shows the output signal for two values of spacing *d*. In the point where the hole and the sensor are precisely aligned, the output is zero regardless of the distance *d* or surrounding temperature.

Figure 7.24 shows another setup, which is useful for measuring angular displacement. The sensor itself is located in the magnetic field produced by two permanent magnets fixed to a rotable frame. The output of the sensor will then be a measure of the rotation of the frame.

Figure 7.25a depicts the use of a single KM110 sensor for detecting rotation and direction of a toothed wheel. The method of direction detection is based on a separate signal processing for the sensor's two half-bridge outputs.

The sensor operates like a magnetic Wheatstone bridge measuring nonsymmetrical magnetic conditions such as when the teeth or pins move in front of the sensor. The mounting of the sensor and the magnet is critical, so the angle between the sensor's symmetry axis and that of the toothed wheel must be kept near zero. Further, both axes (sensor's and wheel's) must coincide. The circuit of Fig. 7.25b connects both bridge outputs to the corresponding amplifiers, and, subsequently, to the low-pass filters and

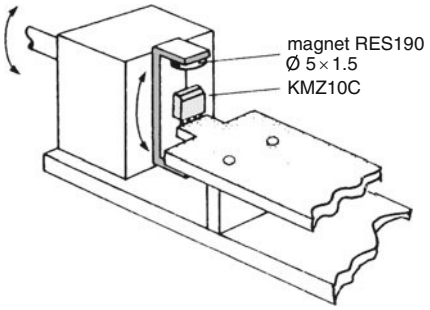

Fig. 7.24 Angular measurement with the KMZ10 sensor

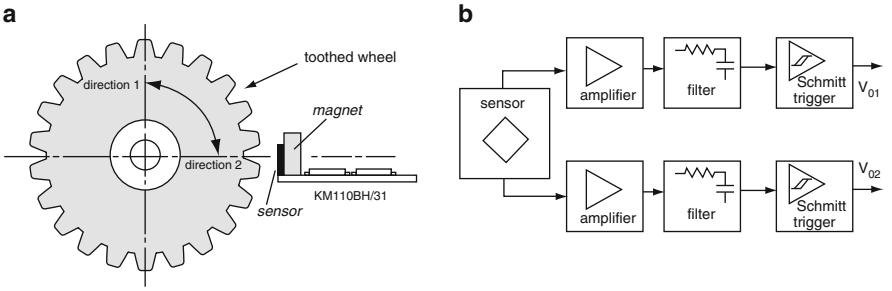

Fig. 7.25 Optimum operating position of a magnetoresistive module (**a**). Note a permanent magnet positioned behind the sensor. Block diagram of the module circuit (**b**)

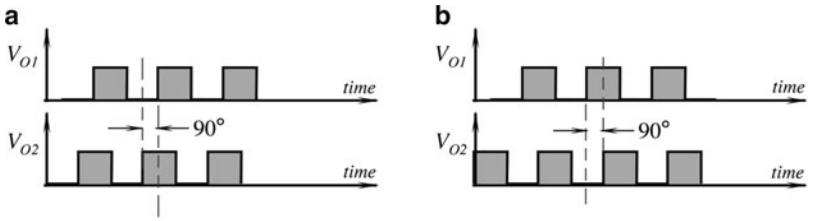

Fig. 7.26 Output signal from the amplifiers for direction 1 (**a**) and 2 (**b**)

Schmitt triggers to form the rectangular output signals. A phase difference between both outputs (Fig. 7.26a, b) is an indication of a rotation direction.

7.3.6 Magnetostrictive Detector

A transducer, which can measure displacement with high resolution across long distances, can be built by using magnetostrictive and ultrasonic technologies [7].

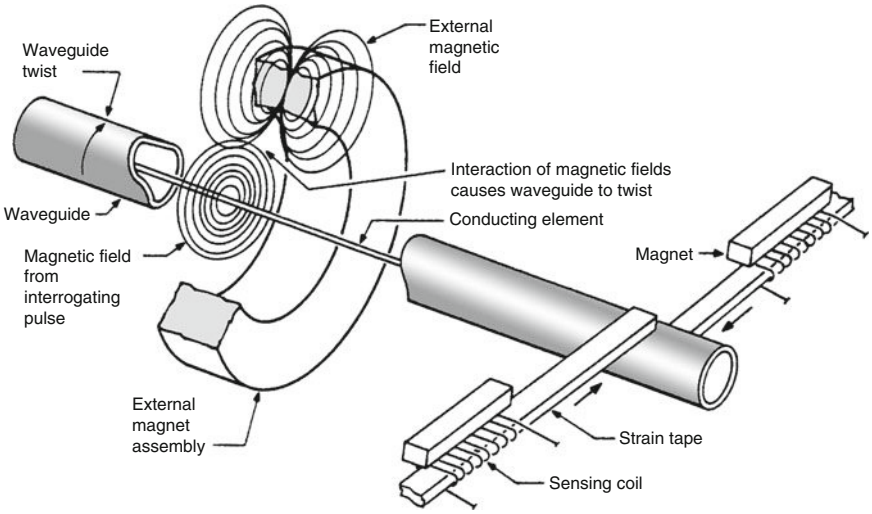

Fig. 7.27 A magnetostrictive detector uses ultrasonic waves to detect position of a permanent magnet

The transducer is comprised of two major parts: a long waveguide (up to 7 m long) and a permanent ring magnet (Fig. 7.27). The magnet can move freely along the waveguide without touching it. A position of that magnet is the stimulus, which is converted by the sensor into an electrical output signal. A waveguide contains a conductor, which upon applying an electrical pulse, sets up a magnetic field over its entire length. Another magnetic field produced by the permanent magnet exists only in its vicinity. Thus two magnetic fields may be setup at the point where the permanent magnet is located. A superposition of two fields results in the net magnetic field, which can be found from the vector summation. This net field, while being helically formed around the waveguide, causes it to experience a minute torsional strain, or twist at the location of the magnet. This twist is known as Wiedemann effect.[5]

Therefore, electric pulses injected into the waveguide's coaxial conductor produce mechanical twist pulses, which propagate along the waveguide with the speed of sound specific for its material. When the pulse arrives at the excitation head of the sensor, the moment of its arrival is precisely measured. One way to detect that pulse is to use a detector that can convert an ultrasonic twitch into an electric output.

[5]Internally, ferromagnetic materials have a structure that is represented by domains, each of which is a region of uniform magnetic polarization. When a magnetic field is applied, the boundaries between the domains shift and the domains rotate, both these effects causing a change in the material's dimensions.

This can be accomplished by piezoelectric sensors, or as shown in Fig. 7.27 by the magnetic reluctance sensor. The sensor consists of two tiny coils positioned near two permanent magnets. The coils are physically coupled to the waveguide and can jerk whenever the waveguide experiences the twitch. This sets up short electric pulses across the coils. Time delay of these pulses from the corresponding excitation pulses in the coaxial conductor is the exact measure of the ring magnet position. An appropriate electronic circuit converts time delay into a digital code representative of a position of the permanent magnet on the waveguide. The advantage of this sensor is in its high linearity (on the order of 0.05% of the full scale), good repeatability (on the order of 3 μm), and a long-term stability. The sensor can withstand aggressive environments, such as high pressure, high temperature, and strong radiation. Another advantage of this sensor is its low temperature sensitivity, which by a careful design, can be achieved on the order of 20 ppm/°C.

Applications of this sensor include hydraulic cylinders, injection molding machines (to measure linear displacement for mold clamp position, injection of molding material and ejection of the molded part), mining (for detection of rocks movements as small as 25 μm), rolling mills, presses, forges, elevators, and other devices where fine resolution along large dimensions is a requirement.

7.4 Optical Sensors

After the mechanical contact and potentionometric sensors, optical sensors are probably the most popular for measuring position and displacement. Their main advantages are simplicity, the absence of a loading effect, and relatively long operating distances. They are insensitive to stray magnetic fields and electrostatic interferences, which makes them quite suitable for many sensitive applications. An optical position sensor usually requires at least three essential components: a light source, a photodetector, and light guidance devices, which may include lenses, mirrors, optical fibers, etc. Examples of single and dual mode fiber optic proximity sensors are shown in Figs. 4.17b and 4.18 (Chap. 4). Similar arrangements are often implemented without the optical fibers, when light is guided toward a target by focusing lenses, and is diverted back to detectors by the reflectors.

7.4.1 Optical Bridge

A concept of a bridge circuit, like a classical Wheatstone bridge, is employed in many sensors and the optical sensor is a good example of such a sensor. One use of the bridge circuit is shown in Fig. 7.28. A four-quadrant photodetector consists of four light detectors connected in a bridge-like circuit. The object must have an optical contrast against the background. Consider a positioning system of a space vehicle (Fig. 7.28a).

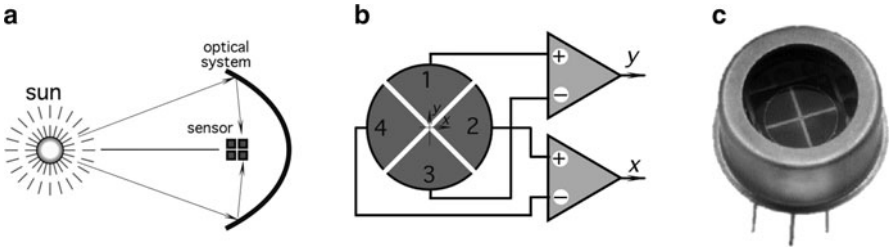

Fig. 7.28 Four-quadrant photodetector. Focusing an object on the sensor (**a**). Connection of the sensing elements to difference amplifiers (**b**). Packaging of the sensor (**c**) (from Advanced Photonix, Inc. Camarillo, CA)

An image of the sun or any other sufficiently bright celestial object is focused by an optical system (e.g. a telescope) on a four-quadrant photodetector. The opposite parts of the detector are connected to the corresponding inputs of the difference amplifiers (Fig. 7.28b). Each amplifier produces the output signal proportional to a displacement of the image from the optical center of the sensor along a corresponding axis. When the image is perfectly centered, both amplifiers produce zero outputs. This may happen only when the optical axis of the telescope passes through the object.

7.4.2 Proximity Detector with Polarized Light

One method of building a better optoelectronic sensor is to use polarized light. Each light photon has specific magnetic and electric field directions perpendicular to each other and to the direction of propagation. The direction of the electric field is the direction of the light polarization (*see* Sect. 3.13.1). Most of the light sources produce light with randomly polarized photons. To make light polarized, it can be directed through a polarizing filter, that is, a special material, which transmits light polarized only in one direction and absorbs and reflects photons with wrong polarizations. However, any direction of polarization can be represented as a geometrical sum of two orthogonal polarizations: one is the same as the filter and the other is nonpassing. Thus, by rotating the polarization of light before the polarizing filter we may gradually change the light intensity at the filter's output (Fig. 7.29).

When polarized light strikes an object, the reflected light may retain its polarization (specular reflection) or the polarization angle may change. The latter is typical for many nonmetallic objects. Thus, to make a sensor nonresponsive to reflective objects (like metal cans, foil wrappers, and the like), it may include two perpendicularly positioned polarizing filters: one at the light source and the other at the detector (Fig. 7.30a, b).

The first filter is positioned at the emitting lens (light source) to polarize the outgoing light. The second filter is at the receiving lens (detector) to allow passage

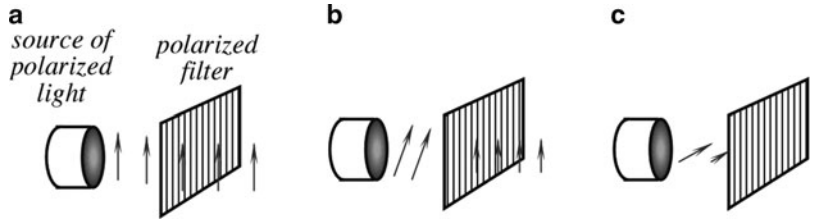

Fig. 7.29 Passing polarized light through a polarizing filter. Direction of polarization is the same as of the filter (**a**). Direction of polarization is rotated with respect to the filter (**b**). Direction of polarization is perpendicular with respect to the filter (**c**)

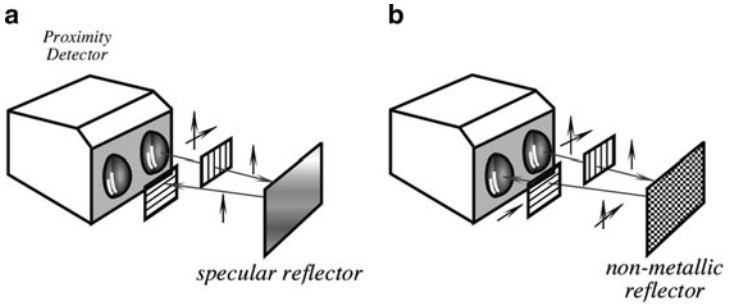

Fig. 7.30 Proximity detector with two polarizing filters positioned at 90° angle with respect to one another. Polarized light returns from the metallic object within the same plane of polarization (**a**); Nonmetallic object depolarizes light, thus allowing it to pass through the polarizing filter (**b**)

of only those components of light, which have a 90° rotation with respect to the outgoing polarization. Whenever light is reflected from a specular reflector (metal) , its polarization direction does not change and the receiving filter will not allow the light to pass to a photodetector. However, when light is reflected in a nonspecular manner, its components will contain a sufficient amount of polarization to go through the receiving filter and activate the detector. Therefore, the use of polarizers reduces false-positive detections of nonmetallic objects.

7.4.3 Fiber-Optic Sensors

Fiber-optic sensors can be used quite effectively as proximity and level detectors. One example of the displacement sensor is shown in Fig. 4.18 (Chap. 4), where intensity of the reflected light is modulated by distance d to the reflective surface.

A liquid level detector (*see* also Sect. 7.7.3) with two fibers and a prism is shown in Fig. 7.31. It utilizes the difference between refractive indices of air (or gaseous phase of a material) and the measured liquid. When the sensor is above the liquid level, a transmitting fiber (on the left) sends most of its light to the receiving fiber

(on the right) due to a total internal reflection in the prism. However, some light rays approaching the prism's reflective surface at angles less than the angle of total internal reflection are lost to the surrounding. When the prism reaches the liquid level, the angle of a total internal reflection changes because the refractive index of a liquid is higher than that of air. This results in much greater loss in the light intensity, which can be detected at the other end of the receiving fiber. The light intensity is converted into an electrical signal by any appropriate light-to-voltage converter. Another version of the sensor is shown in Fig. 7.32, which shows a sensor fabricated by Gems Sensors (Plainville, CT). The fiber is U-shaped and upon being immersed into liquid, modulates the intensity of passing light. The detector has two sensitive regions near the bends where the radius of curvature is the smallest. An entire assembly is packaged into a 5-mm diameter probe and has a repeatability error of about 0.5 mm. Note that the shape of the sensing element draws liquid droplets away from the sensing regions when the probe is elevated above the liquid level.

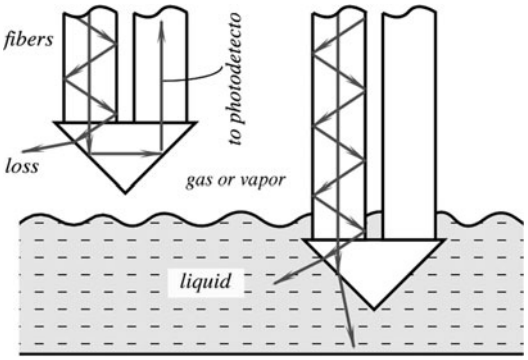

Fig. 7.31 Optical liquid level detector utilizing a change in the refractive index

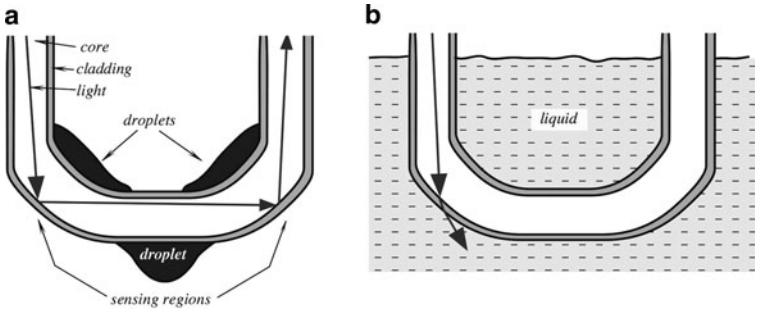

Fig. 7.32 U-shaped fiber optic liquid level sensor. When the sensor is above the liquid level, the light at the output is strongest (**a**); When the sensitive regions touch liquid, the light propagated through the fiber drops (**b**)

7.4.4 Fabry-Perot Sensors

For measuring small displacements with high precision in harsh environment, the so-called Fabry-Perot optical cavity can be employed. The cavity contains two semireflective mirrors facing each other and separated by distance L (Fig. 7.33a). The cavity is injected with light from a known source (a laser, for example) and the photons inside the cavity bounce back and forth between the two mirror interfering with each other in the process. In fact, the cavity is a storage tank for light. At some frequencies of photons, light can pass out of the cavity. A Fabry-Perot interferometer is basically a frequency filter whose transmission frequency is intimately related to the length of the cavity (Fig. 7.33b). As the cavity length changes, the frequencies at which it transmits light change accordingly. If you make one of the mirrors movable, by monitoring the optical transmission frequency, very small changes in the cavity length can be resolved. The narrow bands of transmitted light are separated by frequencies that are inversely proportional to the cavity length:

$$\Delta v = \frac{c}{2L} \tag{7.8}$$

where c is the speed of light. For practical cavities with the mirror separation on the order of 1 μm, typical values of Δv are between 500 MHz and 1 GHz. Thus, by detecting the frequency shift of the transmitted light with respect to a reference light source, changes in the cavity dimensions can be measured with the accuracy comparable with the wavelength of light. Whatever may cause changes in the cavity dimensions (mirror movement), may be the subject of measurements. These include strain, force, pressure, and temperature.

Fabry-Perot cavity-based sensors have been widely used for their versatility; for example, they have been used to sense both pressure and temperature [6–9]. This kind of sensor detects changes in optical path length induced by either a change in the refractive index or a change in physical length of the cavity. Micromachining

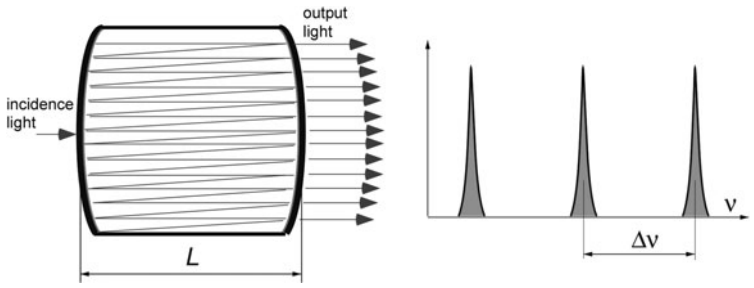

Fig. 7.33 Multiple-ray interference inside Fabry-Perot cavity (**a**). Transmitted frequencies of light (**b**)

techniques make Fabry-Perot sensors more attractive by reducing the size and cost of the sensing element. Another advantage of the miniature Fabry-Perot sensor is that low coherence light sources, such as light-emitting diodes (LEDs) or even light bulbs, can be used to generate the interferometric signal.

A pressure sensor with a Fabry-Perot cavity is shown in Fig. 7.34a. Pressure is applied to the upper membrane. Under pressure, the diaphragm deflects inwardly thus reducing the cavity dimension L. The cavity is monolithically built by the micromachined technology and the mirrors can be either the dielectric layers or metal layers deposited or evaporated during the manufacturing process. The thickness of each layer must be tightly controlled to achieve the target performance of a sensor. An ultraminiature pressure sensor produced by FISO Technologies (www.fiso.com) is shown in Fig. 7.34b. The sensor has very small temperature coefficient of sensitivity (<0.03%) and has an outside diameter of 0.55 mm, which makes it ideal for such critical applications as in the implantable medical devices and other invasive instruments.

A measuring system for the Fabry-Perot sensor is shown in Fig. 7.35. Light from a white light source is coupled through a 2×2 splitter to the optical fiber that in turn is connected to a sensor. The sensor contains a Fabry-Perot interferometer cavity (FPI) and it reflects back light at a wavelength related to the cavity size. The task of the system is to measure the shift in a wavelength. This is accomplished by a white-light cross-correlator that contains a Fabry-Perot wedge. The wedge in effect is a cavity of a linearly variable dimension. Depending of the received wavelength, it passes light only at a specific location of the wedge. The outgoing light position at the wedge may be detected by a position sensitive detector (PSD) that is described in detail in the following pages. The output of the detector directly relates to the input stimulus applied to the FPI sensor.

This method of sensing has advantage of a linear response, insensitivity to the light intensity resulted from the light source or fiber transmission, versatility to

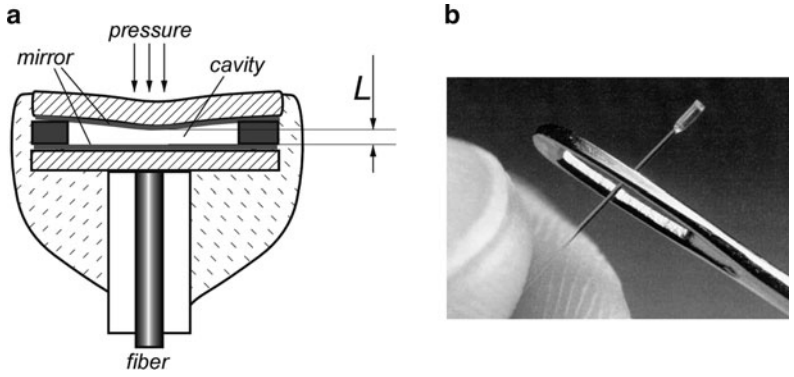

Fig. 7.34 Construction of Fabry-Perot pressure sensor (**a**) and view of FISO FOP-M pressure sensor (**b**)

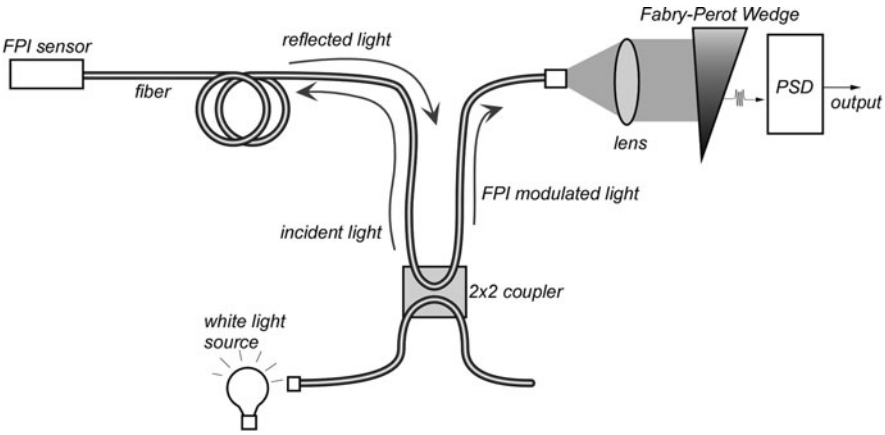

Fig. 7.35 Measuring system for the Fabry-Perot sensor (courtesy of Roctest. www.roctest.com)

measure different stimuli with the same instrument, wide dynamic range (1:15,000), and high resolution. Besides, the fiber optic sensors are immune to many electromagnetic and radiofrequency interferences (EMI and RFI) and can operate reliably in harsh environment without adverse effects. For example, a FPI sensor may function inside a microwave oven.

7.4.5 Grating Sensors

An optical displacement transducer can be fabricated with two overlapping gratings, which serve as a light intensity modulator (Fig. 7.36a). The incoming pilot beam strikes the first, stationary grating, which allows only about 50% of light to pass toward the second, moving grating. When the opaque sectors of the moving grating are precisely aligned with the transmitting sectors of the stationary grating, the light will be completely dimmed out. Therefore, the transmitting light beam intensity can be modulated from 0 to 50% of the pilot beam (Fig. 7.36b). The transmitted beam is focused on a sensitive surface of a photodetector, which converts light into electric current.

The full-scale displacement is equal to the size of an opaque (or clear) sector. There is a trade-off between the dynamic range of the modulator and its sensitivity. That is, for the large pitch of the grating (large sizes of the transparent and opaque sectors), the sensitivity is low, however, the full scale displacement is large. For the higher sensitivity, the grating pitch can be made very small, so that the minute movements of the grating will result in a large output signal. This type of a modulator was used in a sensitive hydrophone [10] to sense displacements of a diaphragm. The grating pitch was 10 μm, which means that the full scale

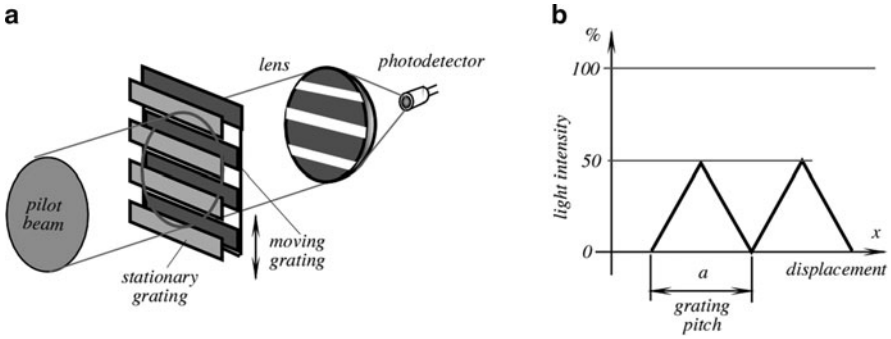

Fig. 7.36 Optical displacement sensor with grating light modulator. Schematic (**a**); transfer function (**b**)

displacement was 5 μm. The light source was a 2-mW He–Ne laser whose light was coupled to the grating through an optical fiber. The tests of the hydrophone demonstrated that the device was sensitive with a dynamic range of 125 dB of pressure as referenced to 1 μPa, with a frequency response up to 1 kHz.

A grating principle of light modulation is employed in very popular rotating or linear encoders, where a moving mask that is usually fabricated in form of a disk has transparent and opaque sections (Fig. 7.37). The photo detector gives a binary output: on and off. In other words, the encoding disk functions as an interrupter of light beams within an optocoupler. When the opaque section of the disk breaks the light beam, the detector is turned off (indicating digital ZERO), and when the light passes through a transparent section, the detector is on (indicating digital ONE). The optical encoders typically employ infrared emitters and detectors operating in the spectral range from 820 to 940 nm. The disks are made of laminated plastic and the opaque lines are produced by a photographic process. Alternatively, the discs are fabricated of metal plates using a photoetching technology.[6] Plastic disks are light, have low inertia, low cost, and excellent resistance to shock and vibration. However, they have a limited operating temperature range. Disks for a broader temperature range are fabricated of etched metal.

There are two types of encoding disks: the incremental, which produces a transient whenever it is rotated for a pitch angle, and the absolute, whose angular position is encoded in a combination of opaque and transparent areas along the radius. The encoding can be based on any convenient digital code. The most common are the gray code, the binary, and the BCD (binary coded decimals).

The incremental encoding systems are more commonly used than the absolute systems, because of their lower cost and complexity, especially in applications, where a displacement (incremental count) is desirable instead of a

[6]Photoetching or photochemical milling parts may be fabricated of a variety materials, including Elgiloy, Nitinol, Titanium, and Kapton® (polyimide film). However, the encoding disks having thickness of 0.005″ typically are etched from stainless steel or beryllium copper alloy.

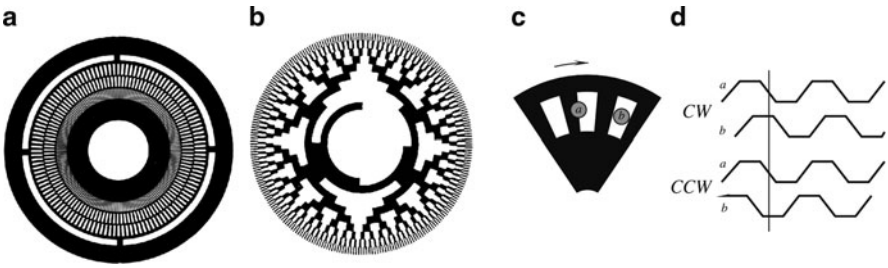

Fig. 7.37 Incremental (**a**) and absolute (**b**) optical encoding disks. When the wheel rotates clockwise (CW), channel *a* signal leads *b* by 90° (**c**); When the wheel rotates counter-clockwise (CCW), channel *b* signal leads *a* by 90° (**d**)

position. When employing the incremental encoding disks, the basic sensing of movement can be made with a single optical channel (an emitter-detector pair), while the speed and incremental position, and direction sensing must use two. The most commonly used approach is a quadrature sensing, where the relative position of the output signals from two optical channels are compared. The comparison provides the direction information, while either of the individual channels gives the transition signal which is used to derive either count or speed information (Fig. 7.37c, d).

7.4.6 Linear Optical Sensors

For precision position measurements over short and long ranges, optical systems operating in the near infrared can be quite effective. An example is a PSD produced for precision position sensing and autofocusing in photographic and video cameras. The position measuring module is of an active type: it incorporates a light emitting diode (LED) and a photodetective PSD. The position of an object is determined by applying the principle of a triangular measurement. Figure 7.38 shows that the near infrared LED through a collimator lens produces a narrow-angle beam (<2°). The beam is a 0.7 ms wide pulse. On striking the object, the beam is reflected back to the detector. The received low intensity light is focused on the sensitive surface of the PSD. The PSD then generates the output signal (currents I_B and I_A), which is proportional to distance x of the light spot on its surface, from the central position. The intensity of a received beam greatly depends on reflective properties of an object. Diffusive reflectivity in the near infrared spectral range is close to that in the visible range, hence, the intensity of the light incident on PSD has a great deal of variations. Nevertheless, accuracy of the measurement depends very little on intensity of the received light.

A PSD operates on the principle of photoeffect. It makes use of a surface resistance of a silicon photodiode. Unlike MOS and CCD sensors integrating multielement photodiode arrays, the PSD has a nondiscrete sensitive area. It provides one-dimensional, or two-dimensional [10] position signals on a light

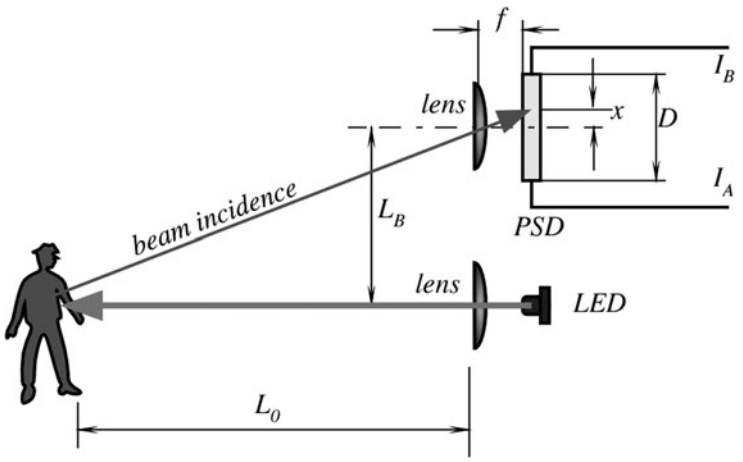

Fig. 7.38 The PSD sensor measures distance by applying a triangular principle

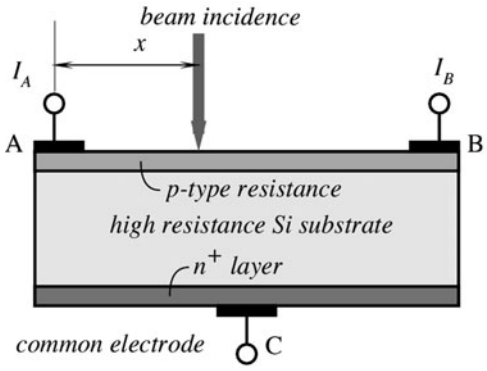

Fig. 7.39 Design of a one-dimensional PSD

spot traveling over its sensitive surface. A sensor is fabricated of a piece of high-resistance silicon with two layers (p and n^+ types) built on its opposite sides (Fig. 7.39). A one-dimensional sensor has two electrodes (A and B) formed on the upper layer to provide electrical contacts to the p-type resistance. There is a common electrode (C) at the center of the bottom layer. Photoelectric effect occurs in the upper pn-junction. The distance between two upper electrodes is D, and the corresponding resistance between these two electrodes is R_D.

Let us assume that the beam incidence strikes the surface at distance x from the A electrode. Then, the corresponding resistance between that electrode and the point of incidence is, respectively, R_x. The photoelectric current I_0 produced by the beam is proportional to its intensity. That current will flow to both outputs (A and B) of the sensors in the corresponding proportions to the

resistances and, therefore, to the distances between the point of incidence and the electrodes

$$I_A = I_0 \frac{R_D - R_x}{R_D} \quad \text{and} \quad I_B = I_0 \frac{R_x}{R_D}. \tag{7.9}$$

.

If the resistances-versus-distances are linear, they can be replaced with the respective distances on the surface

$$I_A = I_0 \frac{D - x}{D} \quad \text{and} \quad I_B = I_0 \frac{x}{D}. \tag{7.10}$$

. To eliminate dependence of the photoelectric current (and of the light intensity), we can use a ratiometric technique, that is we take a ratio of the currents

$$P = \frac{I_A}{I_B} = \frac{D}{x} - 1, \tag{7.11}$$

which we can rewrite for value of x:

$$x = \frac{D}{P + 1}. \tag{7.12}$$

Figure 7.38 shows a geometrical relationships between various distances in the measurement system. Solving two triangles for L_0 yields

$$L_0 = f \frac{L_B}{x}, \tag{7.13}$$

where f is the focal distance of the receiving lens. Substituting (7.12) to (7.13), we arrive at the distance in terms of the current ratio

$$L_0 = f \frac{L_B}{D}(P + 1) = k(P + 1), \tag{7.14}$$

where k is called the module's geometrical constant. Therefore, the distance from the module to the object linearly affects the ratio of the PSD output currents.

A similar operating principle is implemented in an industrial optical displacement sensor (Fig. 7.40a) where PSD is used for measurement small displacements at operating distances of several centimeters. Such optical sensors are highly efficient for the on-line measurements of height of a device (PC-board inspection, liquid and solids level control, laser torch height control, etc.), for measurement of eccentricity of a rotating object, thickness and precision displacement measurements, and detection of presence or absence of an object (medicine bottle caps), etc [11].

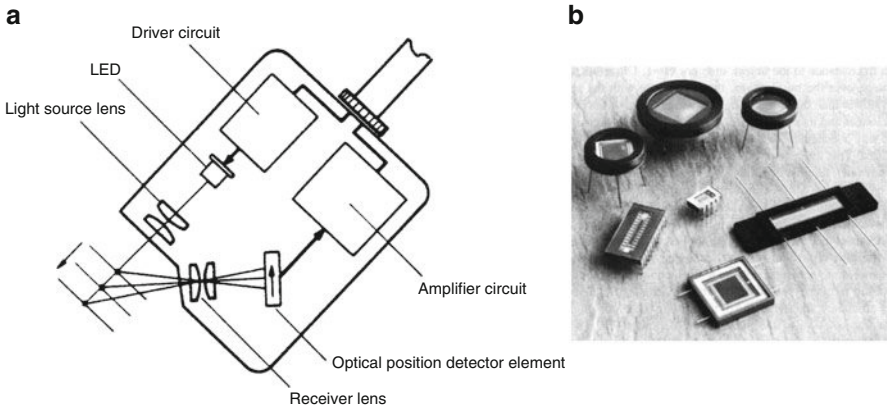

Fig. 7.40 Optical position displacement sensor (**a**) (From Keyence Corp. of America, Fair Lawn, NJ); one- and two-dimensional PSD sensing elements (**b**)

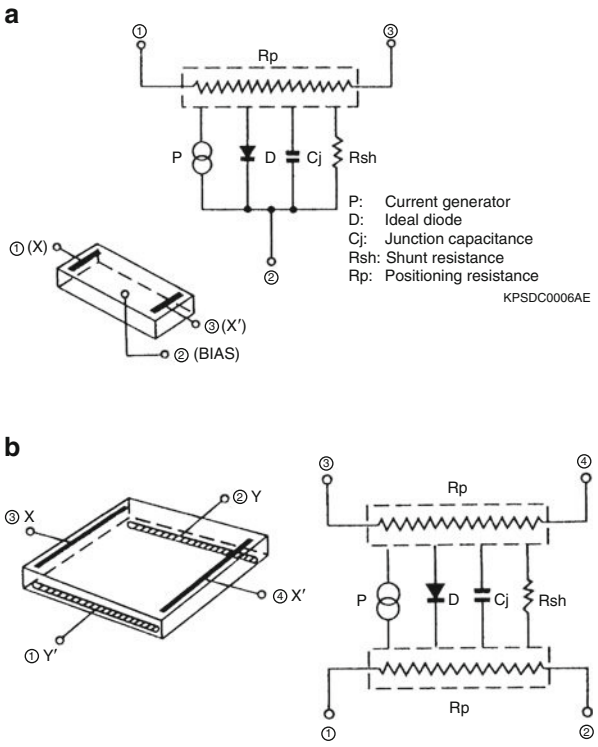

Fig. 7.41 Equivalent circuits for the (**a**) one- and (**b**) two-dimensional position sensitive detectors (Courtesy of Hamamatsu Photonics K.K., Japan)

The PSD elements are produced of two basic types: one- and two-dimensional (Fig. 7.40b). The equivalent circuits of both are shown in Fig. 7.41. Since the equivalent circuit has a distributed capacitance and resistance, the PSD time constant varies depending on the position of the light spot. In response to an input step function, a small area PSD has rise time in the range of 1–2 μs. Its spectral response is approximately from 320 to 1,100 nm, that is, the PSD covers the UV, visible, and near infrared spectral ranges. Small area one-dimensional PSDs have the sensitive surfaces ranging from 1×2 to 1×12 mm, while the large area two-dimensional sensors have square areas with the side ranging from 4 to 27 mm.

7.5 Ultrasonic Sensors

For noncontact distance measurements, an active sensor that transmits some kind of a pilot signal and receives a signal reflected from the object can be designed. The transmitted energy may be in form of any radiation, for instance, electromagnetic in the optical range (like in a PSD which is described above), electromagnetic in the microwave range, acoustic, etc. Transmission and reception of ultrasonic energy is a basis for very popular ultrasonic range meters , and velocity detectors. Ultrasonic waves are mechanical acoustic waves covering frequency range well beyond the capabilities of human ears, i.e. over 20 kHz. However, these frequencies may be quite perceptive by smaller animals, like dogs, cats, rodents, and insects. Indeed, the ultrasonic detectors are the biological ranging devices for bats and dolphins.

When the waves are incident on an object, part of their energy is reflected. In many practical cases, the ultrasonic energy is reflected in a diffuse manner. That is, regardless of the direction where the energy comes from, it is reflected almost uniformly within a wide solid angle, which may approach 180°. If an object moves, the frequency of the reflected wavelength will differ from the transmitted waves. This is called the Doppler effect. [7]

A distance L_0 to the object can be calculated through the speed v of the ultrasonic waves in the media, and the angle, Θ (Fig. 7.42a)

$$L_0 = \frac{vt\cos\Theta}{2}, \tag{7.15}$$

where t is the time for the ultrasonic waves to travel to the object and back to the receiver (thus the denominator 2). If a transmitter and a receiver are positioned

[7] See Sect. 6.2 for the description of the Doppler effect for microwaves. The effect is fully applicable to propagation of any energy having wave nature, including ultrasonic.

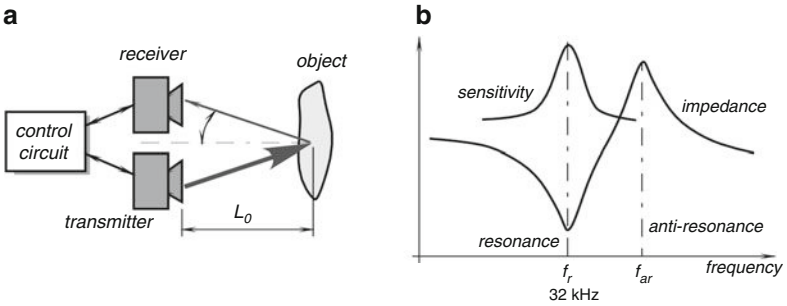

Fig. 7.42 Ultrasonic distance measurement: basic arrangement (**a**); impedance characteristic of a piezoelectric transducer (**b**)

close to each other as compared with the distance to the object, then $\mathrm{Cos}\Theta \approx 1$. Ultrasonic waves have an obvious advantage over the microwaves: they propagate with the speed of sound, which is much slower than the speed of light at what the microwaves propagate. Thus, time t is much longer and its measurement can be accomplished easier and cheaper.

To generate any mechanical waves, including ultrasonic, the movement of a surface is required. This movement creates compression and expansion of medium, which can be gas (air), liquids, or solids.[8] The most common type of the excitation device, which can generate surface movement in the ultrasonic range, is a piezoelectric transducer operating in the so-called motor mode. The name implies that the piezoelectric device directly converts electrical energy into mechanical energy.

Figure 7.43a shows that the input voltage applied to the ceramic element causes it to flex and transmit ultrasonic waves. Because piezoelectricity is a reversible phenomenon, the ceramic generates voltage when incoming ultrasonic waves flex it. In other words, the element may work as both the sonic transmitter and receiver (a microphone). A typical operating frequency of the transmitting piezoelectric element is near 32 kHz. For better efficiency, frequency of the driving oscillator should be adjusted to the resonant frequency f_r of the piezoelectric ceramic (Fig. 7.42b) where the sensitivity and efficiency of the element is the best. When the measurement circuit operates in a pulsed mode, the same piezoelectric element is used for both transmission and reception. When the system requires continues transmission of ultrasonic waves, separate piezoelectric elements are employed for the transmitter and receiver. A typical design of an air-operating sensor is shown in Figs. 7.43b and 7.44a. A directional sensitivity diagram (Fig. 7.44b) is important for a particular application. The narrower the diagram, the more sensitive the transducer.

[8]See Sect. 3.10 for description of sound waves.

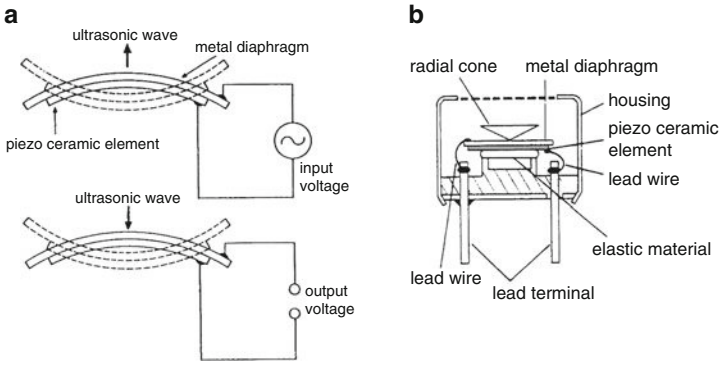

Fig. 7.43 Piezoelectric ultrasonic transducer.
Input voltage flexes the element and transmits ultrasonic waves, while incoming waves produce output voltage (**a**). Open aperture type of ultrasonic transducer for operation in air (**b**) (Courtesy of Nippon Ceramic, Japan)

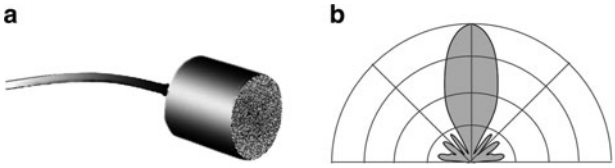

Fig. 7.44 Ultrasonic transducer for air (**a**); directional diagram (**b**)

7.6 Radar Sensors

7.6.1 Micropower Impulse Radar

In 1993, Lawrence Livermore National Laboratory in the United States developed a micropower impulse radar (MIR), which is a low-cost noncontact ranging sensor [20, 21, 23]. The operating principle of the MIR fundamentally is the same as of a conventional pulse radar system, however with several significant differences. The MIR (Fig. 7.45) consists of a white noise generator whose output signal triggers a pulse generator. The pulse generator produces very short pulses with the average rate of 2 MHz $\pm$ 20%. Each pulse has a fixed short duration τ, while the repetition of these pulses is random, according to triggering by the noise generator. The pulses are spaced randomly with respect to one another in a Gaussian noise-like pattern. The distance between the pulses ranges from 200 to 625 ns. It can be said that the pulses have a pulse-position modulation (PPM) by white noise with the maximum index of 20%. In turn, the square wave pulses cause the amplitude modulation (AM)

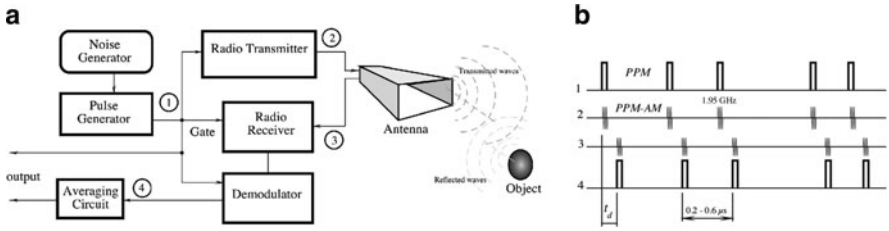

Fig. 7.45 Block-diagram of micropower radar (**a**) and timing diagram (**b**)

of a radio transmitter. The modulation has a 100% depth, that is, the transmitter is turned on and off by the pulses. Such a double-step modulation is called PPM-AM.

The radiotransmitter produces short bursts of high-frequency radio packest, which propagate from the transmitting antenna to the surrounding space. The electromagnetic waves reflect from the objects and propagate back to the radar. The same pulse generator, which modulates the transmitter, with a predetermined delay gates the radio receiver to enable reception by the MIR only during a specific time window when the reflected waves are expected to arrive. Another reason for gating the receiver is to reduce its power consumption. The reflected pulses are received, demodulated (a square-wave shape is restored from the radio-signal), and the time delay with respect to the transmitted pulses is measured, just like in a conventional radar. The time delay is proportional to distance D from the antenna to the object from which the radio waves are reflected: $t_d = 2Dc^{-1}$, where c is the speed of light.

The carrier frequency (center frequency) of the radiotransmitter is either 1.95 or 6.5 GHz. Due to very short modulating pulses, the approximate bandwidth of the radiated signal is very wide – about 500 MHz (for a 1.95 GHz carrier). The spatial distribution of the transmitted energy is determined by the type of antenna. For a dipole antenna it covers nearly 360°, but it may be shaped to a desired pattern by employing a horn, a reflector, or a lens. Because of the unpredictable modulation pattern, the wide bandwidth and a low spectral density of the transmitted signal, the MIR system is quite immune to countermeasures and virtually is stealthy – the radiated energy is perceived by any nonsynchronous receiver as white thermal noise.

The average duty cycle of the transmitted pulses is small (<1%). Since the pulses are spaced randomly, practically any number of identical MIR systems may operate in the same space without a frequency division (that is, they work at the same carrier frequency within the same bandwidth). There is a little chance that bursts from the interfering transmitters overlap, and if they do, the interference level is significantly reduced by the averaging circuit. Nearly 10,000 received pulses are averaged before the time delay is measured.

Another advantage of the MIR are low cost and extremely low power consumption of the radio receiver – about 12 μW. The total power consumption of the entire MIR system is nearly 50 μW. Two AA alkaline batteries may power it continuously for several years.

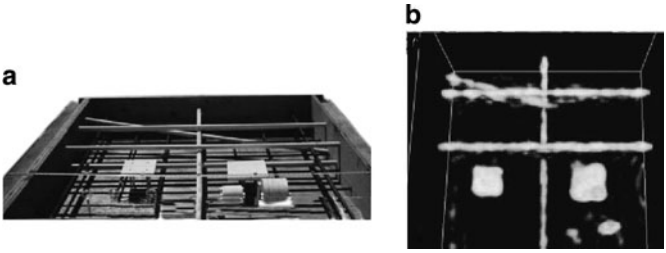

Fig. 7.46 Imaging steel in concrete with MIR.
The internal elements of a concrete slab before pouring (**a**). Reconstructed 3-D MIR image of the elements embedded in the finished, 30-cm-thick concrete slab (**b**)

Applications for the MIR include range meters, intrusion alarms, level detectors, vehicle ranging devices, automation systems, robotics, medical instruments, weapons, novelty products, and even toys where relatively short range of detection is required (Fig. 7.46).

7.6.2 Ground Penetrating Radars

Civil engineering, archeology, forensic science are just few examples of many applications of the high-frequency ground-penetrating radar (GPR). The radar operation is rather classical: it transmits radio waves and receives the reflected signal. The time delay between the transmitted and received signal is the measure of a distance to the reflecting surface. While the radars that operate in air and space have ranges that may reach thousands of kilometers, the GPR range at best is just several hundred meters. A practical GPR operates at frequencies from 500 MHz to 1.5 GHz (www.sensoft.ca). Radio waves do not penetrate far through soils, rocks, and most human-made materials such as concrete. The exponential attenuation coefficient, α, is primarily determined by electrical conductivity of the material. In simple uniform materials, this is usually the dominant factor. In most materials, energy is also lost to scattering from material variability and to water contents. Water has two effects: first, water contains ions, which contribute to bulk conductivity. Second, a water molecule absorbs electromagnetic energy at high frequencies typically above 1 GHz. Figure 7.47 shows that attenuation varies with excitation frequency and material. Thus, practical maximum distance increases for dry materials (Fig. 7.48a). An example of data presented on the radar monitor is shown in Fig. 7.48b.

Lowering frequency improves depth of exploration because attenuation primarily increases with frequency. As frequency decreases, however, two other fundamental aspects of the GPR measurement come into play. First, reducing frequency results in a loss of resolution. Second, if frequency is too low, electromagnetic fields

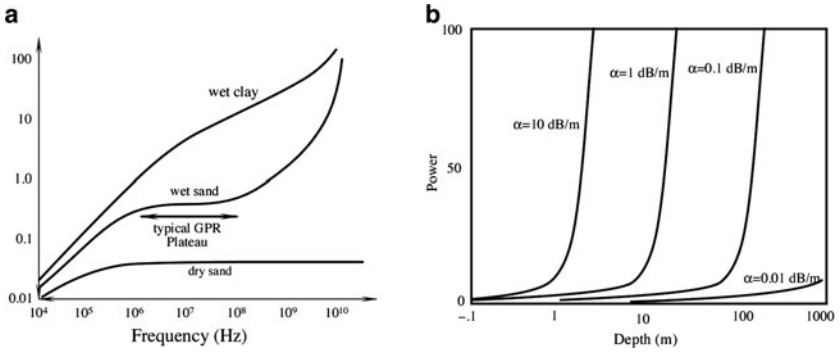

Fig. 7.47 Attenuation of radio waves in different materials (**a**). Attenuation varies with excitation frequency and type of material. At low frequencies (<1 MHz), attenuation is primarily controlled by DC conductivity. At high frequencies (>1 GHz), water is a strong energy absorber. When attenuation limits exploration depth, power must increase exponentially with depth (**b**)

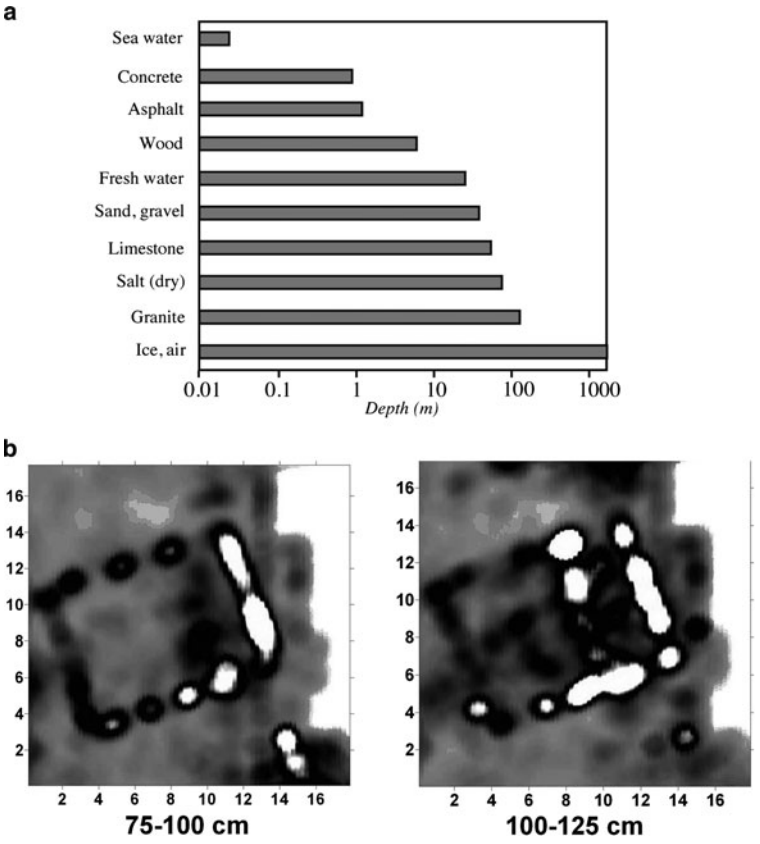

Fig. 7.48 Maximum depth for various materials (**a**). Images of depth slices in a Roman Temple at Petra (Jordan), showing different levels of the temple at different depths (**b**) (courtesy of Prof. L. Conyers, University of Denver)

no longer travel as waves but diffuse, which is the realm of inductive EM or eddy current measurements.

7.7 Thickness and Level Sensors

In many industrial applications, measurement of thickness of a material is essential for manufacturing, process and quality control, safety, airspace, etc. The methods of thickness gauging are ranging from optical to ultrasonic to X-ray. Here we briefly review some less known methods.

7.7.1 Ablation Sensors

Ablation is dissipation of heat by melting and removal of the sacrificial protective layer of a spacecraft during the atmospheric re-entry. Aerospace vehicles subjected to significant aerodynamic heating often rely on ablating thermal protection systems (TPSs) to keep the internal structure and equipment below critical operating temperatures. An ablating TPS undergoes chemical decomposition or phase change (or both) below the internal structure's critical temperature. Incident thermal energy is then channeled into melting, subliming, or decomposing the ablator. Ablator recession rate is directly proportional to the thermal heat flux at the surface. A measure of ablator thickness is required to estimate surface heat flux. Thus an ablation sensor is kind of a position sensor that detects position of the ablation layer's outer surface and provides a measure of the remaining thickness. The ablation sensors can be built into the ablation layer (intrusive sensors) or be noninvasive.

The intrusive sensors include the break-wire ablation gage, radiation transducer (RAT) sensor, and light pipe. The break-wire ablation gage consists of several thin wires implanted at various known levels in an ablator. As the material progressively erodes, each successive wire is broken and results in an open circuit. Figure 7.49a illustrates this concept. In some cases [13] a break-wire doubles as a thermocouple (TC) and each is situated so that no break-wire TC is directly above another. This arrangement allows an unobstructed conduction path through the ablator to each break-wire TC, including those at lower levels. Although the break-wire method provides temperature time histories until the last TC is exposed and destroyed, this method only provides recession data at a few distinct points.

The light pipe sensor consists of quartz fibers implanted in an ablator and terminated at known depths (Fig. 7.49b). When the TPS recedes to where a fiber terminates, light transmits down to a photodiode. This method provides recession data at distinct points only and does not provide temperature data, as the breakwire method does.

Entirely noninvasive sensor for measuring the ablation layer can be build by using a capacitive method. The sensor is made in the form of two electrodes that may have a variety of shapes [12]. The sensor is placed in series with an inductor

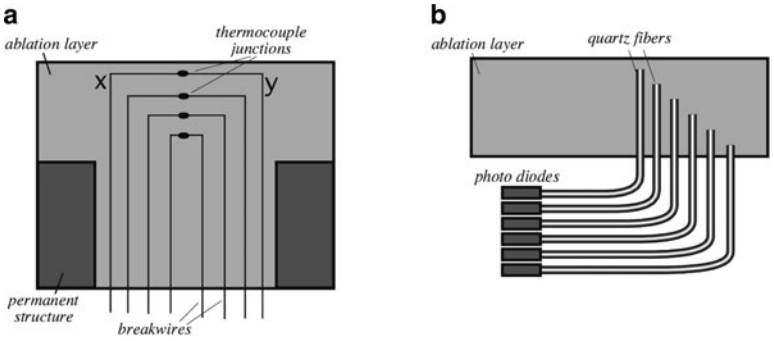

Fig. 7.49 A breakwire concept with thermocouples consisting of metals *x* and *y* (**a**), and a light pipe concept (**b**)

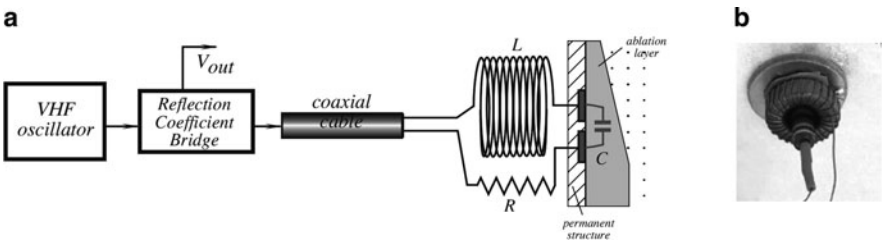

Fig. 7.50 Block diagram of resonant ablation gauge (**a**) and a prototype sensor (**b**)

and a resistor forming a resistive, inductive, and capacitive (RLC) termination to a waveguide (i.e. a coaxial cable). The arrangement shown in Fig. 7.50 is similar to a transmitter-antenna configuration. The RLC termination has a resonant frequency approximated by

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \tag{7.16}$$

When the electromagnetic energy at the resonant frequency is sent down the waveguide, all the energy dissipates in the resistor. If, however, the resonant frequency of the termination changes (say, because of a change in capacitance), a fraction of the energy is reflected back toward the source. As the capacitance continues to change, the energy reflected increases. Antennas that work like this are said to be out of tune. In this situation, one could use a commercially available reflection coefficient bridge (RCB) between the radio frequency (RF) source and the waveguide termination. The RCB generates a DC voltage proportional to the energy reflected. Then the antenna can be adjusted until the bridge output voltage is a minimum and the energy transmitted is a maximum.

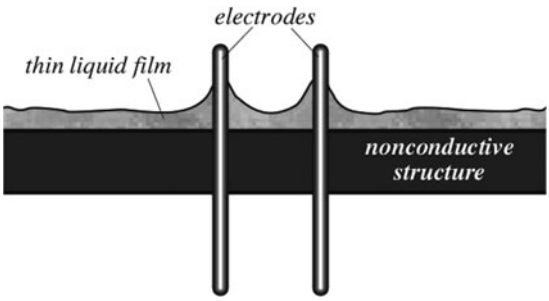

Fig. 7.51 Measurement of thin film liquid by a capacitive method

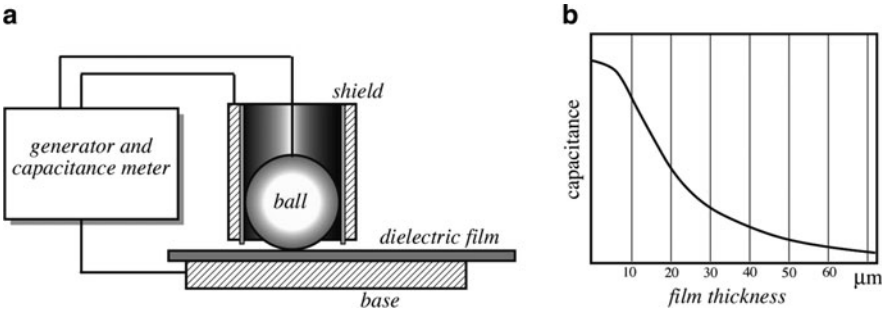

Fig. 7.52 Dry dielectric film capacitive sensor (**a**) and shape of transfer function (**b**). (Adapted from [15])

7.7.2 Thin Film Sensors

Sensors for measuring thickness of a film range from mechanical gauges to optical to electromagnetic and capacitive. Optical methods are limited to transparent or semitransparent films. The planar electrodes that mimic a parallel plate capacitor produce high output, however, to be accurate they and the sampled film must be nearly perfectly parallel, which often is not practical, especially for the surface where the film is positioned on a curvature or moves. Thus, different types of electrodes have been proposed.

Here is an example of a simple capacitive sensor that can measure thickness of liquid film [14]. The liquid film thickness was measured via the capacitance between two small wire probes protruding into the liquid (Fig. 7.51). The liquid acted as a dielectric between two plates of a capacitor with the plates being two small wire probes. Since the liquid has a different dielectric constant than air, a change in liquid level results in a change in the probe's capacitance. The capacitance changes were measured by incorporating the probe into a frequency modulation circuit. A fixed frequency was the input to the circuit, and the output frequency depended on the probe capacitance.

Another type of an electrode is spherical that was proposed for a dry dielectric film [15]. The capacitance is measured between the metal sphere (a stainless steel ball having a diameter between 3 and 4 mm) and a conductive base (Fig. 7.52). To minimize a fringing effect, the ball is surrounded by a driven shield that helps in directing the electric filed only toward the base electrode through the film.

7.7.3 Liquid Level Sensors

There are many ways to detect levels of liquids. They include use of the resistive (*see* Fig. 7.1a), optical (*see* Fig. 7.31), magnetic (*see* Figs. 7.18), and capacitive (*see* Fig. 3.8) sensors. The choice of a particular sensor depends on many factors, but probably the defining factor is a type of a liquid. One of the most challenging are liquid gases, especially liquid helium, which has low density and low dielectric constant, not mentioning its storage in the enclosed Dewar bottles at a cryogenic temperature. In such difficult cases, a transmission line sensor may be quite efficient. A sensor operates on a principle similar the one that was described above for the ablation sensing (Fig. 7.50). For detecting the liquid levels, the transmission line sensor may be constructed as shown in Fig. 7.53.

The probe resembles a capacitive level sensor shown in Fig. 3.8 (Chap. 3), however, its operation does not rely on the liquid dielectric constant as is the case in Fig. 3.8. The probe looks like a long tube with an inner electrode surrounded by the outer cylindrical electrode. The probe is immersed into liquid, which may freely fill the space between the electrodes. The electrodes are fed with a high-frequency signal (about 10 MHz). A length of the probe can be any practical, but for a linear response it is advisable to keep it less than $1/4\lambda$ [16]. The high-frequency signal propagates along the transmission line that is formed by the two electrodes. The liquid fills the space between the electrodes up to a

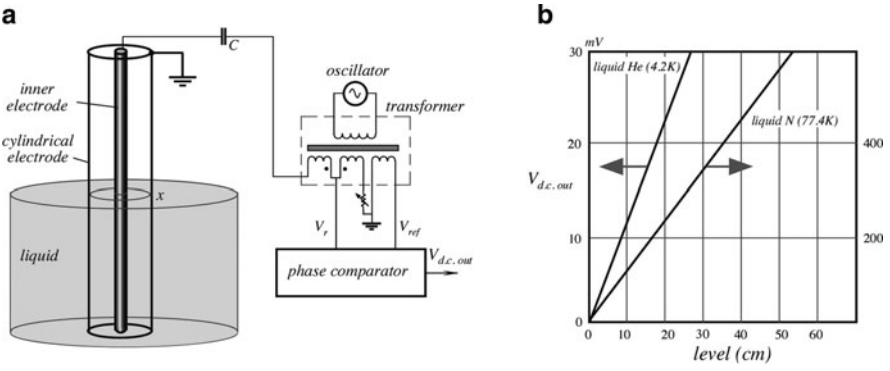

Fig. 7.53 Transmission line probe (**a**) and transfer functions (**b**). (Adapted from [16])

particular level x. Since the dielectric constant of liquid is different from its vapor, the properties of the transmission line depend on the position of the borderline between liquid and vapor, in other words, on the liquid level. The high-frequency signal is partially reflected from the liquid–vapor borderline and propagates back toward the upper portion of the sensor. To some degree, it resembles a radar that sends a pilot signal and received the reflection. By measuring a phase shift between the transmitted and reflected signals, a position of the borderline can be computed. The phase shift measurement is resolved by a phase comparator that produces a d.c. voltage at its output. A higher dielectric constant produces a better reflection and thus sensitivity of the sensor improves accordingly (Fig. 7.53b).

7.8 Pointing Devices

7.8.1 Optical Pointing Devices

Development of personal computers presented a need for another displacement sensor that is called a pointing device. Such device is controlled by a human hand, thus a computer "mouse" was invented. The purpose of the mouse (or a tracking ball) is to move the pointer to a desired x–y coordinate on a computer monitor. A sensor on the pointing device should detect a displacement in a particular direction. The first mice used mechanical rollers coupled to optical encoder disks [similar to that shown in Fig. 7.37a] or electromagnetic pickups. Later, Steve Kirsch invented an optical mouse that required a special reflective pad with a coordinate grid [17]. The newest mice and trackballs employ an optical pickup with illumination by a red

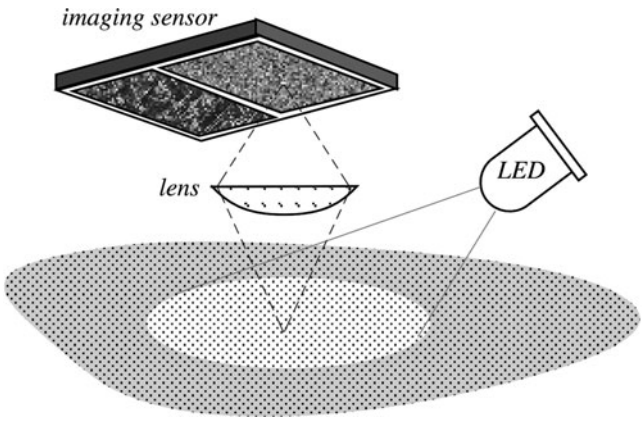

Fig. 7.54 Concept of optical pointing device

(or infrared) LED or laser diode. An optical pointing device contains three essential components: an illuminator, a CMOS optoelectronic image sensor, and a digital signal processing (DSP) chip (Fig. 7.54). The optoelectronic image sensor takes successive pictures of the surface on which the mouse operates, or of a surface pattern on the rotating ball. The DSP chip compares the successive images. Changes between one frame and the next are processed by the image processing part of the chip and translated into movement on the two axes using an optical flow estimation algorithm. For example, the Avago Technologies ADNS-2610 optical mouse sensor processes 1,512 frames per second: each frame consisting of a rectangular array of 18 $\times$ 18 pixels, and each pixel can sense 64 different levels of gray. This advance enabled the mouse to detect relative motion on a wide variety of surfaces, translating the movement of the mouse into the movement of the cursor and eliminating the need for a special mouse pad.

7.8.2 Magnetic Pickup

A Hall sensor can be used as a pickup of a magnetic irregularities build into a ball. An example is a mouse with a ball having the inner magnetic core that contains evenly spaced bumps or depressions [19]. The outer surface of the ball is covered with a nonmagnetic smooth layer for ease of rolling. The magnetic bumps or depressions form irregularities in the magnetic field that are picked up by two x–y Hall sensors and translated into the mouse displacement.

7.8.3 Inertial and Gyroscopic Mice

This device is often called “air mouse” since is does not require a surface to operate. It is the inertial mouse that uses an accelerometer, for example, a tuning fork [18] to detect movement for every axis supported. The user requires only small wrist rotations to move the cursor, reducing user fatigue.

References

1. Barker MJ, Colclough MS (1997) A two-dimensional capacitive position transducer with rotation output. Rev Sci Instrum 68(8):3238–3240
2. Peters RD U.S. Patent No. 5,461,319 Symmetric differential capacitive pressure transducer employing cross coupled conductive plates to form equipotential pairs.
3. De Silva CW (1989) Control sensors and actuators, Prentice Hall, Englewood Cliffs, NJ
4. Linear application handbook (1990) Linear Technology, AN3–9
5. CN-207 Hall Effect IC Applications, Sprague (1986)

6. Halg B (1992) A silicon pressure sensor with a low-cost contactless interferometric optical readout, Sens Actuators A 30:225–229
7. Dakin JP, Wade CA, Withers PB (1987) An optical fiber pressure sensor, SPIE fiber optics '87: fifth international conference on fiber optics and opto-electronics, vol 734, pp 194–201
8. Lee CE and Taylor HF (1991) Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source, J Lightwave Technol 9:129–134
9. Wolthuis RA, Mitchell GL, Saaski E, Hartl JC, Afromowitz MA (1991) Development of medical pressure and temperature sensors employing optical spectrum modulation, IEEE Trans Biomed Eng 38:974–980
10. Spillman WB Jr (1981) Multimode fiber-optic hydrophone based on a schlieren technique. Appl Opt 20:465
11. van Drecht J, Meijer GCM (1991) Concepts for the design of smart sensors and smart signal processors and their applications to PSD displacement transducers. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of technical papers, ©IEEE, pp 475–478
12. Noffz GK, Bowman MP (1996) Design and Laboratory Validation of a Capacitive Sensor for Measuring the Recession of a Thin-Layered Ablator. NASA Technical Memorandum 4777
13. In-Depth Ablative Plug Transducers, (1992) Series #S-2835, Hycal Engineering, 9650 Telstar Avenue, P.O. Box 5488, El Monte, CA
14. Brown RC, Andreussi P, Zanelli S (1978) The use of wire probes for the measurement of liquid film thickness in annular gas-liquid flows, Can J Chem Eng 56:754–757
15. Graham J, Kryzeminski M, Popovic Z (2000) Capacitance based scanner for thickness mapping of thin dielectric films. Rev Sci Intrum 71(5):2219–2223
16. Brusch L, Delfitto G, Mistura G (1999) Level meter for dielectric liquids. Rev Sci Instrum 70(2)
17. Steven Kirsch ST (1985) Detector for electro-optical mouse. U.S. Patent No. 4,546,347, 8 Oct
18. Olson LT (1988) Inertial mouse system. U.S. Patent No. 4,787,051, 22 Nov
19. Solhjell E (1996) Mouse and trackball design with contact-less roller sensor. U.S. Patent No. 5583541, 10 Dec
20. Azevedo SG, Gavel DT, Mast JE, Warhus JP (1995) Landmine detection and imaging using micropower impulse radar (MIR). Proceedings of the workshop on anti-personnel mine detection and removal, 1 July 1995, Lausanne, Switzerland, pp 48–51
21. Dowla FU, Nikoogar F (2007) Multi-pulse multi-delay (MPMD) multiple access modulator for UWB. U.S. Patent No. 7,194,019, 20 Mar
22. Young D et al. (1996) A micromachined variable capacitor for monolithic low-noise VCOs. Solid-state sensor and actuator workshop. Hilton Head, SC
23. McEwan TE (1994) Ultra-wideband radar motion sensor. U.S. Patent No. 5,361,070, 1 Nov

Chapter 8
Velocity and Acceleration

It's a simple task to make a complex system,
It's a complex task to make a simple system

Acceleration is a dynamic characteristic of an object, because according to Newton's second law it essentially requires application of a force. A stationary position does not require an application of a force. A change in a position is associated with velocity and it does not require a force either, unless there is an opposing force, like friction. Acceleration always requires a force. In effect, position, velocity, and acceleration are all related – velocity is a first derivative of a position and acceleration is the second derivative. However, in a noisy environment, taking derivatives may result in extremely high errors, even if complex and sophisticated signal conditioning circuits are employed. Therefore, velocity and acceleration are not derived from the position detectors, but rather measured by special sensors. As a rule of thumb, in low-frequency applications (having a bandwidth on orders from 0 to 10 Hz), position and displacement measurements generally provide good accuracy. In the intermediate-frequency applications (less than 1 kHz), velocity measurement is usually favored. In measuring high-frequency motions with appreciable noise levels, acceleration measurement is preferred.

Velocity (speed or rate of motion) may be linear or angular. That is, it shows how fast an object moves along a straight line or how fast it rotates. The measure of velocity depends on the scale of an object and may be expressed, say, in mm/s or miles/h. Nowadays, speed of a large object, especially of a land or water vehicle, may be very efficiently determined by a geopositioning system (GPS) that operates by receiving radio signals from a number of the Earth satellites and by computing the time delay of signals received from one satellite as compared with the other. When the position of a vehicle is determined with a periodic rate, computation of its velocity is no problem. For smaller objects and shorter distances, GPS is not a solution. Detecting velocity for such objects requires different references. A basic idea behind many sensors for transduction of velocity and acceleration is a measurement of a displacement of an object with respect to some reference object, which in many cases is an integral part of the sensor. *Displacement* is a keyword

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_8, © Springer Science+Business Media, LLC 2010

here. Many velocity or acceleration sensors contain components that are sensitive to a displacement. Thus, position and displacement sensors described in the previous chapter are the integral parts of the velocity sensors and accelerometers. In same instances, however, velocity sensors and accelerometers do not use an intermediate displacement transducer since their motions may be directly converted into electrical signals. For example, moving a magnet through a coil of wire will induce a voltage in the coil according to Faraday's law. This voltage is proportional to the magnet's velocity and the field strength (3.38). Linear velocity transducers use this principle of magnetic induction, with a permanent magnet and a fixed geometry coil, so the output voltage of the coil is directly proportional to the magnet's relative velocity over its working range.

In the velocity sensor, both ends of the magnet are inside the coil. With a single coil, this would give a zero output because the voltage generated by one end of the magnet would cancel the voltage generated by the other end. To overcome this limitation, the coil is divided into two sections.

The north pole of the magnet induces a current in one coil, while the south pole induces a current in the other coil (Fig. 8.1). The two coils are connected in a series-opposite direction to obtain an output proportional to the magnet's velocity. Maximum detectable velocity depends primarily on the input stages of the interface electronic circuit. Minimum detectable velocity depends on the noise floor, and especially of transmitted noise from nearby high AC current equipment. Typical specifications of an electromagnetic sensor are given in Table 8.1. This design is very similar to an LVDT position sensor (Sect. 7.3.1), except that LVDT is an active sensor with a moving ferromagnetic core, while the velocity sensor is a passive device with a moving permanent magnet. That is, this sensor is a current generating device that does not need an excitation signal. Naturally, linear velocity sensors detected velocity along a distance that is limited by the size of the sensor, so

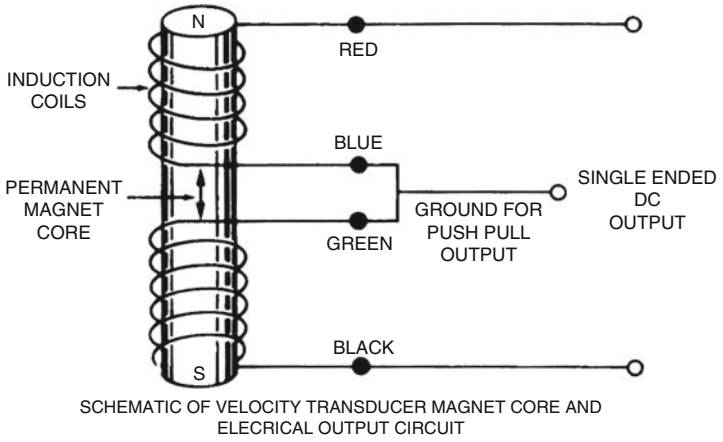

Fig. 8.1 Operating principle of an electromagnetic velocity sensor (Courtesy of Trans-Tek, Inc., Ellington, CT)

Table 8.1 Specification ranges of electromagnetic velocity sensors (from Trans-Tek, Inc., Ellington, CT)

Characteristic	Value
Magnet core displacement (in.)	0.5–24
Sensitivity (mV/in./s)	35–500
Coil resistance (kΩ)	2–45
Coil inductance (henry)	0.06–7.5
Frequency response (Hz) (at load > 100·coil resistance)	500–1,500
Weight (g)	20–1,500

in most cases these sensors measure vibration velocity. An angular version of the same sensor may measure rotation rate continuously for any number of turns.

8.1 Accelerometer Characteristics

Vibration is a dynamic mechanical phenomenon that involves periodic oscillatory motion around a reference position. In some cases (shock analysis, linear acceleration, etc.), the oscillating aspect may be missing, but the measurement and design of the sensor remains the same. An accelerometer can be specified as a single-degree-of-freedom device, which has some type of seismic mass (sometimes called *proof mass*), a spring-like supporting system, and a frame structure with damping properties (Fig. 3.50a).

A mathematical model of an accelerometer is represented by (3.150). To solve the equation, it is convenient to use Laplace transformation, which yields

$$\mathrm{M}s^2X(s) + \mathrm{b}sX(s) + \mathrm{k}X(s) = -\mathrm{M}A(s), \tag{8.1}$$

where $X(s)$ and $A(s)$ are the Laplace transforms of $x(t)$ and $\mathrm{d}^2y/\mathrm{d}t^2$, respectively[1]. Solving the above for $X(s)$ we receive

$$X(s) = -\frac{MA(s)}{Ms^2 + bs + k}. \tag{8.2}$$

We introduce a conventional variable $\omega_0 = \sqrt{k/M}$ and $2\zeta\omega_\mathrm{o} = \mathrm{b/M}$, then (8.2) can be expressed as

$$X(s) = -\frac{A(s)}{s^2 + 2\xi\omega_o s + \omega_o^2}. \tag{8.3}$$

[1] d^2y/dt^2 is the input acceleration of the accelerometer body.

The value of ω_o represents the accelerometer's angular natural frequency, and ζ is the normalized damping coefficient. Let us set

$$G(s) = -\frac{1}{s^2 + 2\xi\omega_o s + \omega_o^2}, \tag{8.4}$$

then, (8.3) becomes: $X(s) = G(s)A(s)$, and its solution can be expressed in terms of the inverse Laplace transform operator as

$$x(\mathrm{t}) = \mathcal{L}^{-1}\{G(s)A(s)\}, \tag{8.5}$$

which from the convolution theorem for the Laplace transform can be expressed as

$$x(t) = \int_0^t g(t-\tau)\alpha(t)d\tau, \tag{8.6}$$

where α is the time-dependent impulse of the accelerometer body and $g(t)$ is the inverse transform $\mathcal{L}^{-1}\{G(s)\}$. If we set $\omega = \omega_0\sqrt{1-\xi^2}$, then (8.6) has two solutions. One is for the underdamped mode ($\xi < 1$),

$$x(t) = \int_0^t -\frac{1}{\omega}e^{-\xi\omega_0(t-\tau)}\sin\omega(t-\tau)\alpha(t)d\tau, \tag{8.7}$$

while for the overdamped mode ($\xi > 1$) the solution is

$$x(t) = \int_0^t -\frac{1}{\omega}\mathrm{e}^{-\xi\omega_0(t-\tau)}\sinh\omega(t-\tau)\alpha(t)d\tau, \tag{8.8}$$

where $\omega = \omega_0\sqrt{\xi^2-1}$. The above solutions can be evaluated for different acceleration inputs applied to the accelerometer base [1].

A correctly designed, installed, and calibrated accelerometer should have one clearly identifiable resonant (natural) frequency and a flat frequency response where the most accurate measurement can be made (Fig. 8.2). Within this flat region, as the vibrating frequency changes, the output of the sensor will correctly reflect the change without multiplying the signal by any variations in the frequency characteristic of the accelerometer. Viscous damping is used in many accelerometers to improve the useful frequency range by limiting effects of the resonant. As a damping medium, silicone oil is used quite often.

When calibrated, several characteristics of an accelerometer should be determined:

1. *Sensitivity* is the ratio of an electrical output to the mechanical input. It is usually expressed in terms of volts per unit of acceleration under the specified conditions. For instance, the sensitivity may be specified as 1 V/g (unit of acceleration: g = 9.80665 m/s^2 at sea level, 45° lat.). The sensitivity is typically

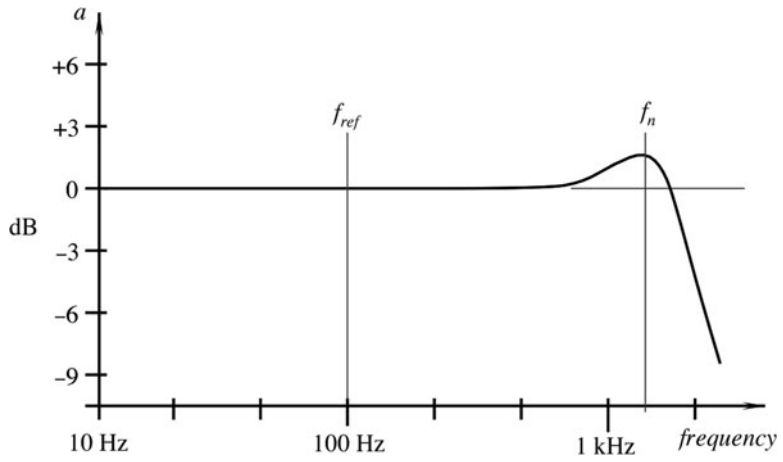

Fig. 8.2 A frequency response of an accelerometer. f_n is a natural frequency; f_{ref} is the reference frequency

measured at a single reference frequency of a sine-wave shape. In the USA, it is 100 Hz, while in most European countries it is 160 Hz[2].

2. *Frequency response* is the outputs signal over a range of frequencies where the sensor should be operating. It is specified with respect to a reference frequency which is where the sensitivity is specified.
3. *Resonant frequency* in an undamped sensor shows as a clearly defined peak that can be 3–4 dB higher than the response at the reference frequency. In a near-critically damped device the resonant may not be clearly visible; therefore, the phase shift is measured. At the resonant frequency, it is 180° of that at the reference frequency.
4. *Zero stimulus output* (for the capacitive and piezoresistive sensors) is specified for the position of the sensor where its sensitive (active) axis is perpendicular to Earth's gravity. That is, in the sensors that have a DC component in the output signal, the gravitational effect should be eliminated before the output as no mechanical input is determined.
5. *Linearity* of the accelerometer is specified over the dynamic range of the input signals.

When specifying an accelerometer for a particular application, one should answer a number of questions, such as the following:

1. What is the anticipated magnitude of vibration or linear acceleration?
2. What is the operating temperature and how fast the ambient temperature may change?

[2]These frequencies are chosen because they are removed from the power line frequencies and their harmonics.

3. What is the anticipated frequency range?
4. What linearity and accuracy are required?
5. What is the maximum tolerable size?
6. What kind of power supply is available?
7. Are any corrosive chemicals or high moisture present?
8. What is an anticipated overshock?
9. Are intense acoustic, electromagnetic, or electrostatic fields present?
10. Is the machinery grounded?

8.2 Capacitive Accelerometers

An accelerometer requires a special relatively massive component, whose movement lags behind that of the accelerometer's housing, which is coupled to the object under study. Then, a displacement transducer can be employed to generate an electrical signal as function of the acceleration. This massive component is usually called either a seismic, inertial, or proof mass. No matter what the sensors' design or the conversion technique is, an ultimate goal of the measurement is the detection of the mass displacement with respect to the accelerometer housing. Hence, any suitable displacement transducer capable of measuring microscopic movements under strong vibrations or linear acceleration can be used in an accelerometer. A capacitive displacement conversion is one of the proven and reliable methods. A capacitive acceleration sensor essentially contains at least two components, where the first is a "stationary" plate (i.e., connected to the housing) and the other is a plate attached to the inertial mass, which is free to move inside the housing. These plates form a capacitor whose value is function of a distance d between the plates (3.23). It is said that the capacitor value is modulated by the acceleration. A maximum displacement, which is measured by the capacitive accelerometer, rarely exceeds 20 μm. Hence, such a small displacement requires a reliable compensation of drifts and various interferences. This is usually accomplished by use of a differential technique, where an additional capacitor is formed in the same structure. The value of the second capacitor must be close to that of the first, and it should be subjected to changes with a 180° phase shift. Then, acceleration can be represented by a difference in values between the two capacitors.

Figure 8.3a shows a cross-sectional diagram of a capacitive accelerometer where an internal mass is sandwiched between the upper cap and the base [2]. The mass is supported by four silicon springs (Fig. 8.3b). The upper plate and the base are separated from it by respective distances d_1 and d_2. All three parts are micromachined from a silicon wafer. Figure 8.4 is a simplified circuit diagram for a capacitance-to-voltage converter, which in many respects is similar to the circuit of Fig. 5.32.

A parallel plate capacitor C_{mc} between the mass and the cap electrodes has a plate area S_1. The plate spacing d_1 can be reduced by an amount Δ when the mass moves toward the upper plate. A second capacitor C_{mb} having a different plate area

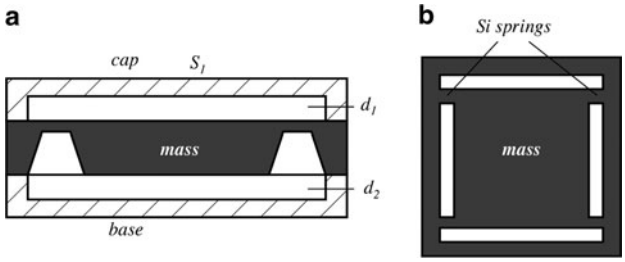

Fig. 8.3 Capacitive accelerometer with a differential capacitor side cross-sectional view (**a**); top view of a seismic mass supported by four silicon springs (**b**)

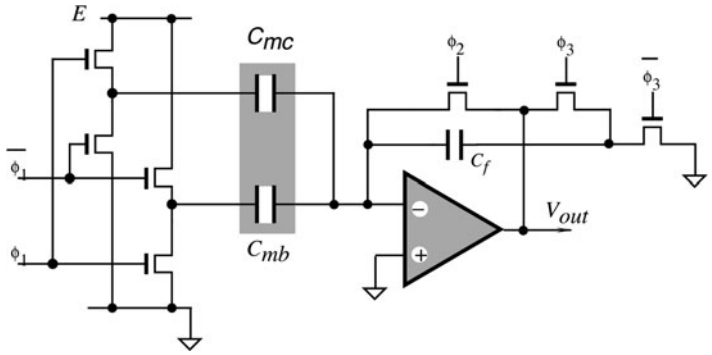

Fig. 8.4 Circuit diagram of a capacitance-to-voltage conversion suitable for an integration on silicon

S_2 appears between the mass and the base. When mass moves toward the upper plate and away from the base, the spacing d_2 increases by Δ. The value of Δ is equal to the mechanical force F_m acting on the mass divided by the spring constant k of the silicon springs:

$$\Delta = \frac{F_{\mathrm{m}}}{k}. \tag{8.9}$$

Strictly speaking, the accelerometer equivalent circuit is valid only when electrostatic forces do not affect the mass position, that is, when the capacitors depend linearly on F_m [3]. When an accelerometer serves as the input capacitor to a switched-capacitor summing amplifier, the output voltage depends on value of capacitors, and subsequently on force

$$V_{out} = 2E\frac{C_{mc} - C_{mb}}{C_f}. \tag{8.10}$$

The above equation is true for small changes in sensor's capacitances. The accelerometer output is also a function of temperature and a capacitive mismatch. It is advisable that it is calibrated over an entire temperature range and an appropriate correction is made during the signal processing. Another effective method of assuring high stability is to design self-calibrating systems, which makes use of electrostatic forces appearing in the accelerometer assembly when high voltage is applied to either a cap or a base electrode.

8.3 Piezoresistive Accelerometers

As a sensing element, a piezoresistive accelerometer incorporates a strain gauge that measures strain in the mass-supporting springs. The strain can be directly correlated with the magnitude and rate of the mass displacement and, subsequently, with an acceleration. These devices can sense accelerations within a broad frequency range: from near DC up to 13 kHz. With a proper design, they can withstand overshock up to 10,000g. Naturally, a dynamic range (span) is somewhat narrower (1,000g with error less than 1%). The overshock is a critical specification for many applications. Piezoresistive accelerometers with discrete, epoxy-bonded strain gauges tend to have undesirable output temperature coefficients. Since they are manufactured separately, the gauges require individual thermal testing and parameter matching. This difficulty is virtually eliminated in modern sensors that use micromachining technology of the silicon wafers.

An example of a wide dynamic range solid-state accelerometer is shown in Fig. 8.5. It was developed by Endevco/Allied Signal Aerospace Co. (Sunnyvale, CA). The microsensor is fabricated from three layers of silicon. The inner layer or

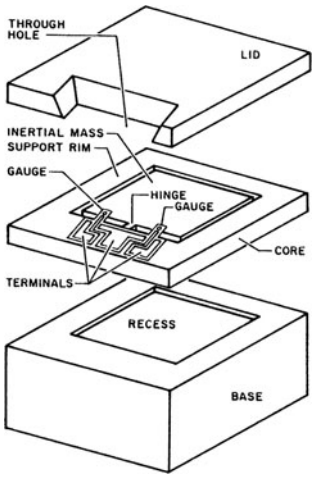

Fig. 8.5 Exposed view of a piezoresistive accelerometer

the core consists of an inertial mass, and the elastic hinge. The mass is suspended inside an etched rim on the hinge, which on either side has piezoresistive gauges. The gauges detect motion about the hinge. The outer two layers, the base and the lid, protect the moving parts from the external contamination. Both parts have recesses to allow the inertial mass to move freely [4]. Several important features are incorporated into the sensor. One is that the sensitive axis lies in the plane of the silicon wafer, as opposed to many other designs where the axis is perpendicular to the wafer. A mechanical integrity and reliability are assured by the fabrication of all components of the sensor from a single silicon crystal.

When acceleration is applied along the sensitive axis, the inertial mass rotates around the hinge. The gauges on both sides of the hinge allow rotation of the mass to create compressive stress on one gauge and tensile on the other. Since gauges are very short, even a small displacement produces large resistance changes. To trim the zero balance of the piezoresistive bridge, there are five trimming resistors positioned on the same crystal (not shown in the figure).

8.4 Piezoelectric Accelerometers

The piezoelectric effect (do not confuse it with a piezoresistive effect) has a natural application in sensing vibration and acceleration. The effect is a direct conversion of mechanical energy into electrical energy (Sect. 3.6) in a crystalline material composed of electrical dipoles. These sensors operate from frequency as low as 2 Hz and up to about 5 kHz, they posses good off-axis noise rejection, high linearity, and a wide operating temperature range (up to 120°C). While quartz crystals are occasionally used as sensing elements, the most popular are ceramic piezoelectric materials, such as barium titanate, lead zirconite titanate (PZT), and lead metaniobite. A crystal is sandwiched between the case and the seismic mass that exerts on it the force proportional to the acceleration (Fig. 8.6). In miniature

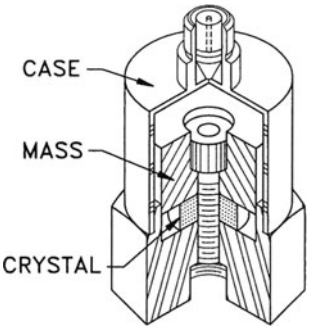

Fig. 8.6 A basic schematic representation of a piezoelectric accelerometer. Acceleration of the case moves it relative to the mass, which exerts a force on the crystal. The output is directly proportional to the acceleration or vibration level

sensors, a silicon structure is usually employed. Since silicon does not possess piezoelectric properties, a thin film of lead titanate can be deposited on a micromachined silicon cantilever to fabricate an integral miniature sensor. For good frequency characteristics, a piezoelectric signal is amplified by a charge-to-voltage, or current-to-voltage converter, which usually is built into the same housing as the piezoelectric crystal.

8.5 Thermal Accelerometers

8.5.1 Heated Plate Accelerometer

Since the basic idea behind an accelerometer is a measurement of movement of a seismic mass, a fundamental formula of heat transfer can be used for that measurement (see (3.148)). A thermal accelerometer, as any other accelerometer, contains a seismic mass that is suspended by a thin cantilever and positioned in close proximity with a heat sink, or between two heat sinks (Fig. 8.7) [5]. The mass and the cantilever structure are fabricated using a micromachined technology. The space between these components is filled with a thermally conductive gas. The mass is heated by a surface-deposited or imbedded heater to a defined temperature T_1. Under the no-acceleration conditions, a thermal equilibrium is established between the mass and the heat sinks: the amount of heat q_1 and q_2 conducted to the heat sinks through gas from the mass is a function of distances M_1 and M_2.

The temperature at any point in the cantilever beam supporting the seismic mass[3] depends on its distance from the support x and the gaps at the heat sinks. It can be found from

$$\frac{d^2T}{dx^2} - \lambda^2 T = 0, \tag{8.11}$$

where

$$\lambda = \sqrt{\frac{K_g(M_1 + M_2)}{L_{si}DM_1M_2}}, \tag{8.12}$$

[3]Here we assume steady-state conditions and neglect radiative and convective heat transfers.

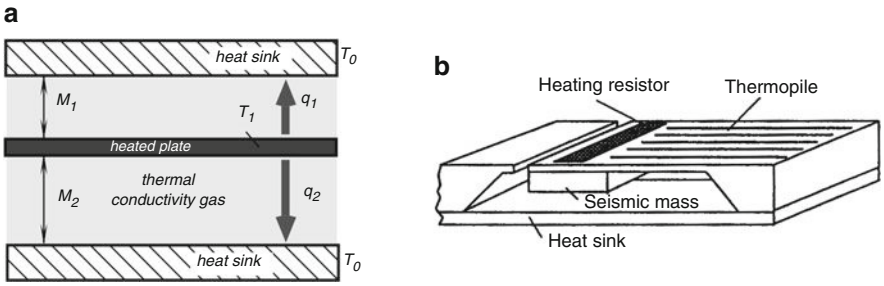

Fig. 8.7 Thermal accelerometer.
(**a**) A cross-section of the heated part; (**b**) an accelerometer design shown without the roof (adapted from [5])

where K_g and K_{si} being thermal conductivities of gas and silicon respectively, and D is the thickness of a cantilever beam. For boundary conditions, where the heat sink temperature is 0, a solution of the above equation for the temperature of the beam is

$$T(x) = \frac{P \sinh(\lambda x)}{WDK_{si}\lambda \cosh(\lambda L)}, \tag{8.13}$$

where W and L are the width and length of the beam, and P is the thermal power. To measure that temperature, a temperature sensor can be deposited on the beam. It can be done by integrating silicone diodes into the beam[4], or by forming serially connected thermocouples (a thermopile) on the beam surface. Eventually, the measured beam temperature in form of an electrical signal is a measure of acceleration. The sensitivity of a thermal accelerometer (about 1% of change in the output signal per g) is somewhat smaller than that of the capacitive or piezoelectric types; however, it is much less susceptible to such interferences as ambient temperature or electromagnetic and electrostatic noise.

8.5.2 *Heated Gas Accelerometer*

Another interesting accelerometer uses gas as a seismic mass. The heated gas accelerometer (HGA) was developed by MEMSIC Corporation (www.memsic.com). It is fabricated on a micromachined CMOS chip and is a complete biaxial motion measurement system. The principle of operation of the device is based on heat transfer by forced convection. As it was described in Chap. 3, heat can be transferred by conduction, convection, and radiation. Convection can be natural (caused by gravity) or forced (by applying an artificial external force, like that produced by a blower). In HGA, such force is produced by acceleration. The sensor measures the internal changes in heat transfer of the trapped gas. The sensor is functionally

[4]See Chap. 16 for a description of a Si diode as a temperature sensor.

equivalent to traditional inertial mass accelerometers. The inertial mass in the sensor is gas that is thermally nonhomogeneous. The gaseous inertial mass provides some advantages over use of the traditional solid inertial mass. The most important advantage is a shock survival up to 50,000g leading to significantly lower failure rates.

The sensor contains a micromachined plate adjacent to a sealed cavity filled with gas. The plate has an etched cavity (trench). A single heat source, centered in the silicon chip is suspended across the trench (Fig. 8.8). Equally spaced are four temperature sensors that are aluminum/polysilicon thermopiles (TP) – serially connected thermocouples. The TPs are located equidistantly on all four sides of the heat source (dual axis). Note that a TP measures only a temperature gradient so that the left and right TP in fact is a single TP where the left portion is the location of "cold" junctions and the right portion is for the "hot" junctions (see Sect. 16.4 for the operating principle of a thermocouple). A thermopile instead of a thermocouple is used for a sole purpose – to increase the electrical output signal. Another pair of the junctions is for measuring a thermal gradient along the y axis.

Under zero acceleration, a temperature distribution across the gas cavity is symmetrical about the heat source, so that the temperature is the same at all four TP junctions, causing each pair to output zero voltage. The heater is warmed to a temperature that is well above ambient and typically is near 200°C. Figure 8.8a shows two thermopile junctions (TP) for sensing a temperature gradient along a single axis. Gas is heated so that it is hottest near the heater and rapidly cools down toward the left and right temperature sensors (thermopile junctions). When no force acts on gas, temperature has a symmetrical cone-like distribution around the heater, where temperatures T_1 at the left TP is equal to temperature T_2 of the right TP.

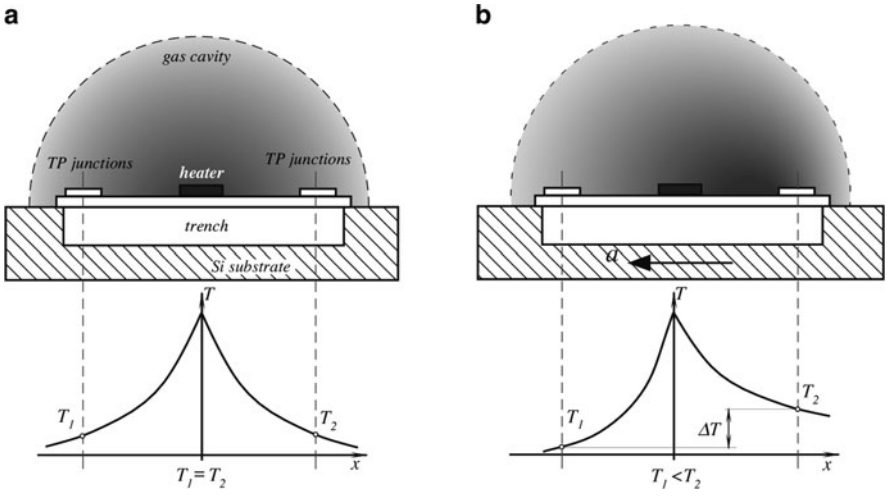

Fig. 8.8 Cross-sectional view of the HGA sensor along x axis (**a**). Heated gas is symmetrical around the heater (**b**). Acceleration causes heated gas to shift to the right, resulting in temperature gradient

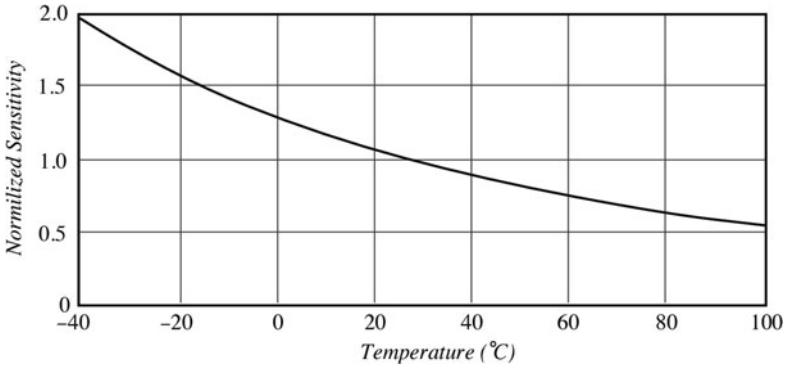

Fig. 8.9 Thermal accelerometer (HGA) sensitivity to ambient temperature

Acceleration in any direction will disturb the temperature profile, due to a convection heat transfer, causing it to be asymmetrical. Figure 8.8b shows acceleration a in a direction of the arrow. Under the acceleration force, warm gaseous molecules shift toward the right TP and transfer to it a portion of their thermal energy. The temperature, and hence voltage output of the opposite TP junctions will then be different so that $T_1 < T_2$. The differential temperature ΔT and thus voltage at the thermopile outputs becomes directly proportional to the acceleration. There are two identical acceleration signal paths on the device, one to measure acceleration in the x axis and one to measure acceleration in the y axis.

The HGA is capable of measuring accelerations with a full-scale range from below $\pm 1.0g$ to above $\pm 100g$. It can measure both the dynamic acceleration (e.g., vibration) and static acceleration (e.g., gravity). The analog output voltages from the chip are available in absolute and ratiometric modes. The absolute output voltage is independent of the supply voltage, while the ratiometric output voltage is proportional to the supply voltage. The typical noise floor is below 1 mg/Hz, allowing sub-milli-g signals to be measured at very low frequencies. The frequency response, or the capability to measure fast changes in acceleration is defined by design. A typical -3 dB rolloff occurs at above 30 Hz but is expandable with a compensation to over 160 Hz

It should be noted that for the HGA sensor, the output sensitivity changes with ambient temperature. The sensitivity change is shown in Fig. 8.9. To compensate for the change, an imbedded temperature sensor (RTD or silicon junction may be provided on the chip as a compensating temperature sensor).

8.6 Gyroscopes

Before advancement of a GPS (global positioning system), besides a magnetic compass, a gyroscope probably was the most common navigation sensor. In many cases where a geomagnetic field was either absent (in Space), or is altered by

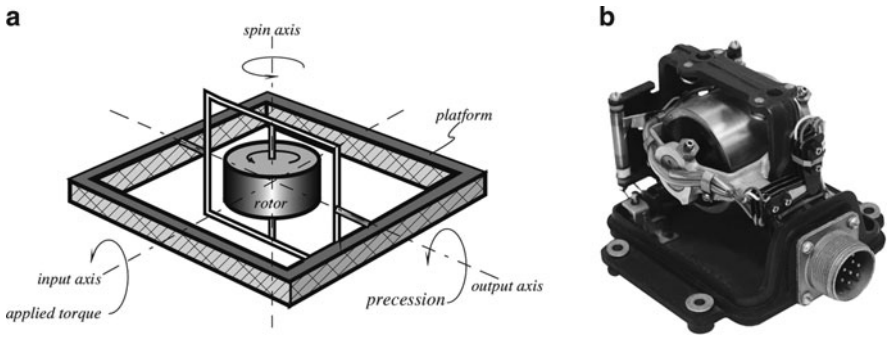

Fig. 8.10 Mechanical gyroscope with a single degree-of-freedom (**a**) and early auto-pilot gyroscope (**b**)

presence of some disturbances, a gyroscope was an indispensable sensor for defining the position of a vehicle, such as an aircraft or missile (Fig. 8.10b). Nowadays, applications of gyroscopes are much broader than for navigation. They are used in the stabilization devices, weapons, robotics, tunnel mining, and in many other systems where a stable directional reference is required.

The earliest known gyroscope containing a rotating massive sphere was made in 1817 by German Johann Bohnenberger. In 1832, American Walter R. Johnson developed a gyroscope that was based on a rotating disk. In 1852, Léon Foucault used a rotating disc in an experiment involving the rotation of the Earth. Foucault coined the name *gyroscope* from the Greek word *skopeein* – to see and *gyros* – circle or rotation.

A gyroscope, or a *gyro* for short, is a "keeper of direction", like a pendulum in a clock is a "keeper of time". A gyro operation is based on the fundamental principle of the conservation of angular momentum: in any system of particles, the total angular momentum of the system relative to any point fixed in space remains constant, provided no external forces act on the system.

8.6.1 Rotor Gyroscope

A mechanical gyro is comprised of a massive disk free to rotate about a spin axis (Fig. 8.10a), which itself is confined within a framework that is free to rotate about one or two axes. Hence, depending on the number of rotating axes, gyros can be either of a single or two-degree-of-freedom types. The two qualities of a gyro account for it usefulness are: (1) the spin axis of a free gyroscope will remain fixed with respect to space, provided there are no external forces to act upon it and (2) a gyro can be made to deliver a torque (or output signal) that is proportional to the angular velocity about an axis perpendicular to the spin axis.

When the wheel (rotor) freely rotates, it tends to preserve its axial position. If the gyro platform rotates around the input axis, the gyro will develop a torque around a perpendicular (output) axis, thus turning its spin axis around the output axis. This phenomenon is called precession of a gyro. It can be explained by Newton's law of motion for rotation: the time rate of change of angular momentum about any given axis is equal to the torque applied about the given axis. That is, when a torque T is applied about the input axis, and the speed ω of the wheel is held constant, the angular momentum of the rotor may be changed only by rotating the projection of the spin axis with respect to the input axis. In other words, the rate of rotation of the spin axis about the output axis is proportional to the applied torque

$$T = I\omega\Omega, \tag{8.14}$$

where Ω is the angular velocity about the output axis and I is the inertia of a gyro wheel about the spin axis. To determine the direction of precession, the following rule can be used: precession is always in such a direction as to align the direction of rotation of the wheel with the direction of rotation of the applied torque.

The accuracy of mechanical gyros greatly depends on the effects that may cause additional unwanted torques and cause drifts. The sources of these are friction, imbalanced rotor, magnetic effects, etc. One method that is widely used to minimize rotor friction is to eliminate the suspension entirely by floating the rotor and the driving motor in a viscous, high-density liquid, such as one of the fluorocarbons. This method requires close temperature control of the liquid and also may suffer from aging effects. The other method of friction reduction is to use the so-called gas bearings, where the shaft of the rotor is supported by high pressure helium, hydrogen or air. And even a better solution is to support the rotor in vacuum by an electric field (electrostatic gyros). A magnetic gyro consists of a rotor supported by a magnetic field. In that case, the system is cryogenically cooled to temperatures where the rotor becomes super conductive. Then, an external magnetic field produces enough counterfield inside the rotor that the rotor floats in a vacuum. These magnetic gyroscopes also are called cryogenic.

8.6.2 Monolithic Silicon Gyroscopes

While a spinning-rotor gyroscope for many years was the only practical choice, its operating principle really does not lend itself to design of a small monolithic sensor that is required by many modern applications. Conventional spinning rotor gyroscopes contain parts such as gimbals, support bearings, motors, and rotors that need accurate machining and assembly; these aspects of construction prohibit conventional mechanical gyroscopes from ever becoming a low-cost device. Wear on the motors and bearings during operation means that the gyroscope will only meet the performance specifications for a set number of running hours. Other methods for sensing direction and velocity of motion have been developed. Often,

a GPS would be the ideal choice. Yet, frequently it just cannot be employed in Space, under water, in tunnels, inside buildings, or whenever the size and cost are of paramount importance. Use of MEMS micromachined technology allows designing of a miniature gyroscope where the rotating disk is replaced with a vibrating element. The design takes advantage of the techniques developed in the electronic industry and is highly suited to high-volume manufacture. Besides, the vibrating gyro is much more robust and can withstand the environments typical of many military and aerospace applications.

All vibrating gyroscopes rely on the phenomenon of the Coriolis acceleration. Coriolis effect is an inertial force described by the nineteenth-century French engineer-mathematician Gustave-Gaspard Coriolis in 1835. Coriolis showed that, if the ordinary Newtonian laws of motion of bodies are to be used in a rotating frame of reference, an inertial force – acting to the right of the direction of body motion for counterclockwise rotation of the reference frame or to the left for clockwise rotation – must be included in the equations of motion. The Coriolis acceleration of a body appears, whenever that body moves linearly in a frame of reference that is rotating about an axis perpendicular to that of the linear motion. The resulting acceleration, which is directly proportional to the rate of turn, occurs in the third axis that is perpendicular to the plane containing the other two axis (Fig. 8.12a). In a micromachined gyro, the rotation is replaced by vibration and the resulting acceleration can be detected and related to the rate of motion. Instead of a mass following a circular trajectory as for the conventional spinning-rotor gyroscope, the mass can be suspended and made to move linearly in simple harmonic motion.

There are several practical ways to build a vibrating gyro, however, all of them can be divided into three principle groups [6]:

1. Simple oscillators (mass on a string, beams)
2. Balanced oscillators (tuning forks)
3. Shell resonators (wine glass, cylinder, ring)

All three categories have been implemented in the actual designs.

One of the first such devices was a two-gimbal structure supported by torsional flexures (Fig. 8.11). It is undercut and free to move in the active area. In operation, the outer, or "motor" is driven at constant amplitude by electrostatic torquing using electrodes placed in close proximity. This oscillatory motion is transferred to the inner gimbal along the stiff axis of the inner flexures, setting up an oscillating momentum vector with the inertial element. In the presence of an angular rotational rate normal to the plane of the device, the Coriolis force will cause the inner gimbal to oscillate about its weak axis with a frequency equal to the drive frequency and with an amplitude proportional to the inertial input rate. Maximum resolution is obtained when the outer gimbal is driven at a resonant frequency of the inner gimbal. The readout of the output motion is accomplished by setting the differential change in capacitance between the inner gimbal and a pair of electrodes. When operated open loop, the angular displacement of the inner gimbal about the output axis is proportional to the input rate. That is, the output angle Θ is proportional to an

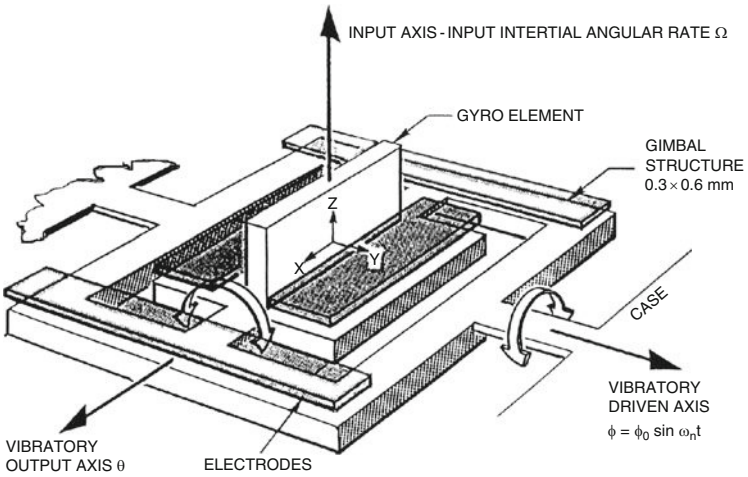

Fig. 8.11 Vibratory rate gyro concept (from [7])

inertia ration term, the drive angle, ϕ_0, the mechanical Q, and the input rate Ω. It is inversely proportional to the drive frequency ω_n

$$\Theta = \left[\frac{\mathbf{I_x} + \mathbf{I}_y - \mathbf{I}_z}{\mathbf{I_x}}\right] \frac{\phi_0 \Omega Q}{\omega_n}. \tag{8.15}$$

In practical application, the device is operated closed loop and the inner gimbal is rebalanced to null in phase and in quadrature. A detailed description of the gyroscope may be found elsewhere [7].

A more recent design that also belongs to the third category was developed by British Aerospace Systems and Equipment along with its partner Sumitomo Precision Products Company Ltd [8].

The design is based on a ring resonator that is micromachined in silicon. Silicon has remarkable mechanical properties (see Sect. 18.1.1 for details), specifically, in its crystalline state silicon has a fracture limit of 7 GPa, which is higher than the majority of steels. Coupled with this is a low density of 2,330 kg/m^3, resulting in a very robust material under its own weight. The gyro resonator is etched out of the crystalline silicon material. This ensures that the properties of the resonator are stable over life and environment. The planar vibrating ring structure has all the vibration energy in one plane. As such, under angular rate, there is no coupling of vibration from one crystal plane to another, so that the vibrating parameters are very stable over temperature.

In order for the resonator to function correctly, it must be supported in a way that allows it to vibrate as freely as possible. The sensing element is shown in Fig. 8.12b. The resonator comprises a 6 mm silicon ring, supported by eight radially compliant spokes, which are anchored to a support frame 10 by 10 mm. Current carrying

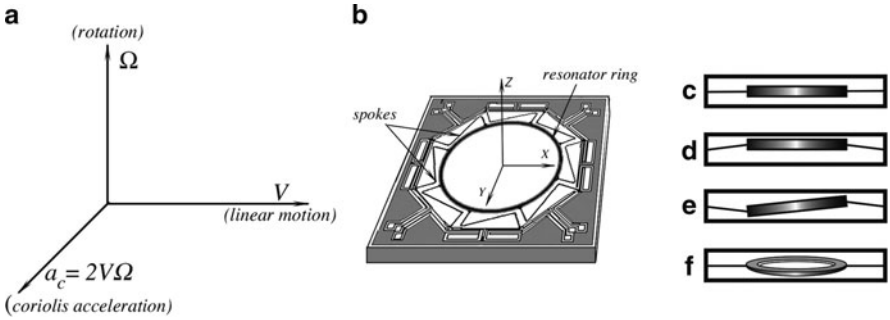

Fig. 8.12 Coriolis acceleration (**a**). Vibrating ring micromachined structure (**b**); effects of acceleration on the vibrating modes of the ring (**c–f**)

conductors are deposited and patterned onto the top surface only, and pads for wire bonding are located on the outer support frame. The chip is anodically bonded to a supporting glass structure, which is thermally matched to the silicon. There are eight identical conducting loops, which each follow the pattern, bond pad – along length of support leg – around 1/8 segment of ring – along length of next support leg – bond pad. Each leg thus contains two conductors, one each from adjacent loops, in addition to a third conductor, which lies between them, to minimize capacitive coupling. The resonator may be excited into vibration by any suitable transducers. These may function by means of optical, thermal expansion, piezoelectric, electrostatic, or electromagnetic effects, for example. The excitation may be applied to the support structure that carries resonator, or directly to the resonator itself. The fundamental vibration mode is at 14.5 kHz. Figures 8.12c–f shows the effects of linear and angular acceleration on the resonator. Figure 8.12c shows a side view of the resonator under conditions of no acceleration, Fig. 8.12d shows effect of the z axis linear acceleration, Fig. 8.12e shows the effect of angular acceleration about the x axis, Fig. 8.12f shows the effect of angular acceleration about the y axis. Since the ring position changes with respect to the frame, what is required is a combination of displacement pickup transducers to detect a particular movement of the resonator. The resonator vibration may, for example, be sensed by transducers working electromagnetically, capacitively, optically, piezoelectrically, or by means of strain gauges. In this particular design, a magnetic pickup is employed by pattern conductive loops along with a magnetic field, which is perpendicular to the plane of the ring. The magnetic field is provided by a Samarium Cobalt and the entire structure is housed in a standard hermetic metal IC can package.

8.6.3 Optical (Laser) Gyroscopes

Modern development of sensors for guidance and control applications are based on employing the so-called Sagnac effect, which is illustrated in Fig. 8.13 [9]. Two

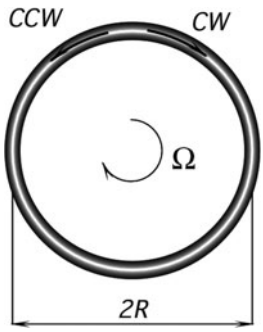

Fig. 8.13 Sagnac effect

beams of light generated by a laser propagate in opposite directions within an optical ring having refractive index n and radius R. One beam goes in clockwise (CW) direction, while the other in a counterclockwise (CCW) direction. The amount of time that light takes to travel within the ring is $\Delta t = 2\pi R/nc$, where c is the speed of light. Now, let us assume that the ring rotates with angular rate Ω in the clockwise direction. In that case, light will travel different paths at two directions. The CW beam will travel $l_{cw} = 2\pi R + \Omega R\Delta t$, while the CCW beam will travel $l_{ccw} = 2\pi R - \Omega R\Delta t$. Hence, the difference between the paths is

$$\Delta l = \frac{4\pi\Omega R^2}{nc}. \tag{8.16}$$

Therefore, to accurately measure Ω, a technique must be developed to determine Δl. There are three basic methods are known for the path detection: (1) optical resonators, (2) open-loop interferometers, and (3) closed-loop interferometers.

For the ring laser gyro, measurements of Δl are made by taking advantages of the lasing characteristics of an optical cavity (that is, of its ability to produce coherent light). For lasing to occur in a closed optical cavity, there must be an integral number of wavelengths about the complete ring. The light beams that do not satisfy this condition, interfere with themselves as they make subsequent travel about the optical path. In order to compensate for a change in the perimeter due to rotation, the wavelength λ and frequency ν of the light must change

$$-\frac{dv}{v} = \frac{d\lambda}{\lambda} = \frac{dl}{l}. \tag{8.17}$$

The above is a fundamental equation relating frequency, wavelength, and perimeter change in the ring laser. If the ring laser rotates at a rate Ω, then (8.16) indicates that light waves stretch in one direction and compress in the other direction to meet the criteria for the lasing of an integral number of wavelengths about the ring. This, in turn, results in a net frequency difference between the light beams. If the two beams are bit together (mixed), the resulting signal has frequency

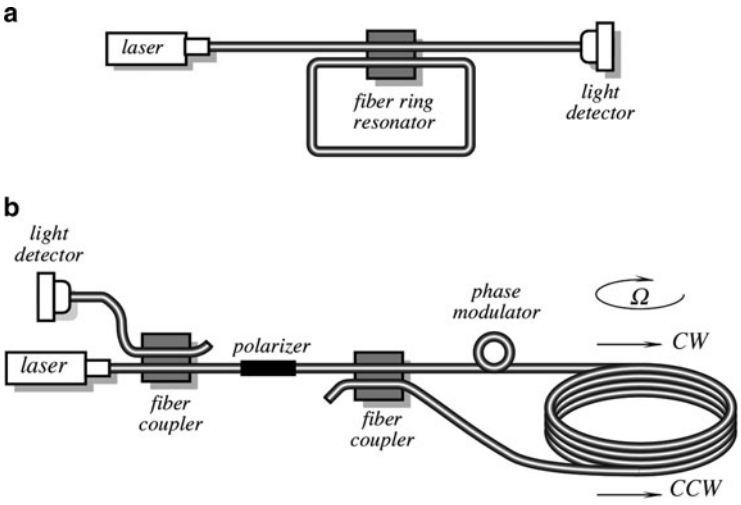

Fig. 8.14 Fiber optic ring resonator (**a**); fiber optic analog coil gyro (**b**) (adapted from [9])

$$F = \frac{4A\Omega}{\lambda nl}, \tag{8.18}$$

where A is the area enclosed by the ring.

In practice, optic gyros are designed with either a fiber ring resonator, or the fiber coil where the ring has many turns of the optical fiber [10]. The optic ring resonator is shown in Fig. 8.14a It consists of a fiber loop formed by a fiber beam splitter that has a very low cross-coupling ratio. When the incoming beam is at the resonant frequency of the fiber ring, the light couples into the fiber cavity and the intensity in the exiting light drops. The coil fiber gyro (Fig. 8.14b) contains a light source and the detector coupled to the fiber. The light polarizer is positioned between the detector and the second coupler to insure that both counter-propagating beams traverse the same path in the fiber optic coil [11]. The two beams mix and impinge onto the detector, which monitors the cosinusoidal intensity changes caused by rotationally induced phase changes between the beams. This type of optical gyro provides a relatively low cost, small-size rotation-sensitive sensor with a dynamic range up to 10,000. Applications include yaw and pitch measurements, attitude stabilization and gyrocompassing. A major advantage of optical gyros is their ability to operate under hostile environments that would be difficult, if not impossible, for the mechanical gyros.

8.7 Piezoelectric Cables

A piezoelectric effect is employed in a vibration sensor built with a mineral insulated cable. Such a cable generates an electric signal in its internal conductor when the outer surface of the cable is subjected to variable compressions. The

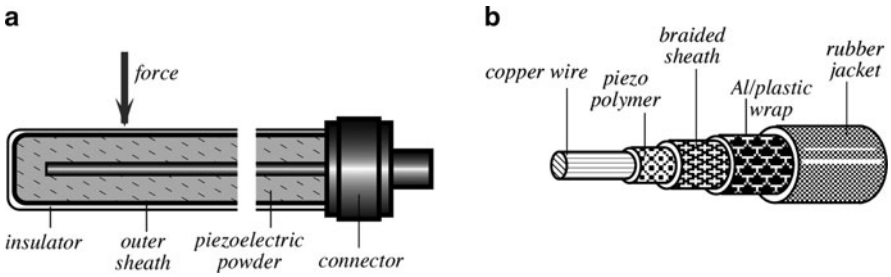

Fig. 8.15 Piezoelectric cable sensors. Construction of *Vibracoax* (**a**); polymer film as a voltage generating component (**b**) (adapted from [13])

piezoelectric Vibracoax™ cables[5] have been used in various experiments to monitor the vibration in compressor blades in turboshaft aircraft engines. Other applications include detection of insects in silos and automobile traffic analysis. In these applications, the cables are buried in the highway pavement, positioned perpendicular to the traffic. When properly installed, they last for at least 5 years [12]. The sensors are designed to be sensitive primarily to vertical forces. A piezoelectric cable consists of a solid insulated copper sheath having 3 mm outer diameter, piezoelectric ceramic powder (see Sect. 3.6), and an inner copper core (Fig. 8.15a). The powder is tightly compressed between the outer sheath and the core. Usually, the cable is welded at one end and connected to a 50Ω extension cable at the other end.

Another method of fabrication of the piezoelectric cables is to use a PVDF polymer film as a component in the cable insulation (Fig. 8.15b). The PVDF can be made piezoelectric, thus giving cable-sensing properties. When a mechanical force is applied to the cable, the piezoelectric film is stressed, which results in the development of electric charges of the opposite polarities on it surfaces. The inner copper wire and the braided sheath serve as charge pickup electrodes.

For the cable to possess piezoelectric properties, its sensing component (the ceramic powder or polymer film) must be poled during the manufacturing process [14]. That is, the cable is warmed up to near the Curie temperature and subjected to high voltage to orient ceramic dipoles in the powder, or polymer dipoles in the film, then cooled down while the high voltage is maintained. When the cable sensor is installed into the pavement (Fig. 8.16), its response should be calibrated, because the shape of the signal and its amplitude depend not only on the properties of the cable, but also on the type of the pavement and subgrade.

The electrical output is proportional to the stress imparted to the cable. The long, thin piezoelectric insulating layer provides a relatively low output impedance (600 pF/m), unusual for a piezoelectric device. The dynamic range of the cable is substantial (>200 dB), sensing distant small amplitude vibrations caused by rain or

[5] www.irdinc.com.

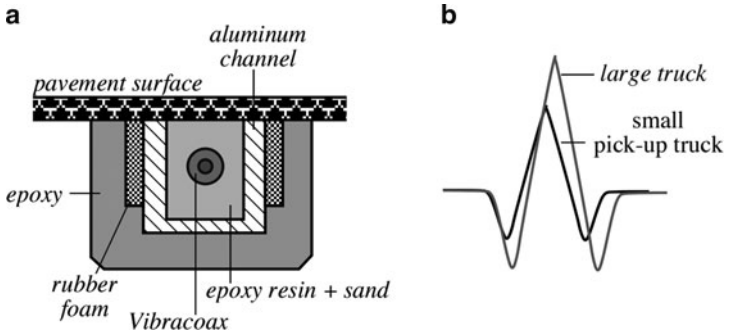

Fig. 8.16 Application of the piezoelectric cables in highway monitoring. Sensor installation in the pavement (**a**); shape of electrical response (**b**)

Table 8.2 Typical properties of a piezoelectric cable (from [15])

Parameter	Units	Value
Capacitance at 1 kHz	pF/m	600
Tensile strength	MPa	60
Young's modulus	GPa	2.3
Density	kg/m^3	1,890
Acoustic impedance	MRayl	4.0
Relative permittivity	At 1 kHz	9
Tan δ_e	At 1 kHz	0.017
Hydrostatic piezo coefficient	pC/N	15
Longitudinal piezo coefficient	Vm/N	250×10^{-3}
Hydrostatic piezo coefficient	Vm/N	150×10^{-3}
Electromechanical coupling	%	20
Energy output	mJ/strain (%)	10
Voltage output	kV/strain (%)	5

hail, yet responding linearly to the impacts of heavy trucks. The cables have withstood pressures of 100 MPa. The typical operating temperature range is −40 to +125°C. Table 8.2 lists typical properties of a piezoelectric cable.

8.8 Gravitational Sensors

The central assumption of the Einstein's general relativity is the equivalence principle, which states that gravity is a force that arises from being in an accelerated reference frame. It can be said then that the force resulted from acceleration and the force arisen from gravity are essentially the same thing. Thus, according to the Einstein's theory, if we want to detect direction and magnitude of a gravitational force, an accelerometer is our sensor. However, if only direction of the gravitational force is of interest, a special type of an accelerometer should be considered. A response of such an accelerometer is not function of the force vector

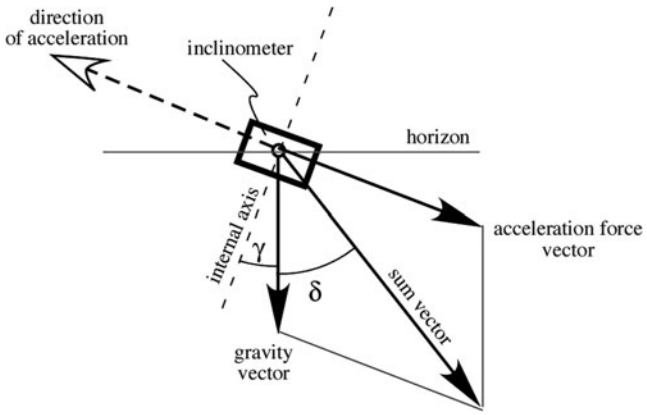

Fig. 8.17 Inclination sensor (inclinometer) positioned on an accelerating vehicle

magnitude but only of the force direction. Special types of accelerometers that are responsive to direction of acceleration (gravity) are called inclinometers or tilt detectors. Such a detector has its own internal axis with respect to which the direction is measured. A response of the detector is an electric signal representative of an angle between the internal axis and the gravity vector. It should be noted that if an inclinometer is positioned on an accelerating vehicle it will produce an erroneous signal because the sensor responds to the sum of two vectors – earth gravity and acceleration as shown in Fig. 8.17. In a stationary or steady moving vehicle, an inclinometer will measure angle α between its own internal axis and the gravity vector. If the vehicle carrying the inclinometer accelerates, an acceleration force will appear in the opposite direction from the direction of motion. The two vectors, gravity and acceleration, will sum and the inclinometer will respond with an error equal to angle δ. Naturally, if a purpose of an inclinometer is to detect direction of acceleration, this will be not an error but a useful output. Inclinometers are available for one or two axes of tilt.

Inclination detectors are employed in ground and air-based vehicles, road construction, machine tools, inertial navigation systems, handheld video monitors (to control orientation of an image), robots, electronic games, and other applications requiring a gravity reference or direction of acceleration. An old and still quite popular detector is a mercury switch (Fig. 8.18a, b). The switch is made of a nonconductive (often glass) tube having two electrical contacts and a drop of mercury. When the sensor is positioned with respect to the gravity force in such a way as the mercury moves away from the contacts, the switch is open. A change in the switch orientation causes the mercury to move to the contacts and touch both of them, thus closing the switch. One popular application of this design is in a household thermostat, where the mercury switch is mounted on a bimetal coil (see Fig. 3.38), which serves as an ambient temperature sensor. Winding or unwinding the coil in response to the room temperature affects the switch orientation.

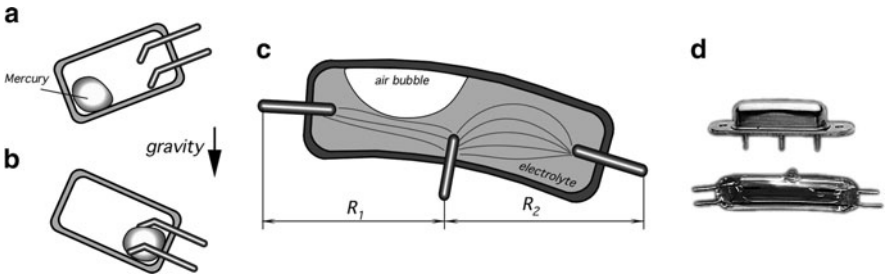

Fig. 8.18 Conductive gravitational sensors.
Mercury switch in the open position (**a**) and closed (**b**) positions; electrolytic tilt sensor (**c**) and electrolytic sensor housings (**d**)

Opening and closing the switch controls a heating/cooling system. An obvious limitation of this design is its on–off operation (a "bang-bang" controller in the engineering jargon), which is an equivalent of a proportional control system. A mercury switch is a threshold device, which snaps when its rotation angle exceeds a predetermined value.

To measure angular displacement with higher resolution, a more complex sensor is required. One elegant design is shown in Fig. 8.18c. It is called the *electrolytic tilt sensor*. A small slightly curved glass tube is filled with partly conductive electrolyte. Three electrodes are built into the tube: two at the ends and the third electrode is at the center of the tube. An air bubble resides in the tube and may move along its length as the tube tilts. Electrical resistances between the center electrode and each of the end electrodes depend on the position of the bubble. As the tube shifts away from the balance position, the resistances increase or decrease proportionally. The electrodes are connected into a bridge circuit that is excited with an AC current to avoid damage to the electrolyte and electrodes. The electrolytic tilt sensors are available[6] in different designs (Fig. 8.18d) for a wide spectrum of angular ranges from ± 0.5 to $\pm 80^\circ$. Correspondingly, the shapes of the glass tubes vary from slightly curved to doughnut-like.

A more advanced inclination sensor employs an array of photodetectors [16]. The detector is useful in civil and mechanical engineering for the measurements of shapes of complex objects with high resolution. Examples include the measurement of ground and road shapes, and the flatness of an iron plate, which cannot be done by the conventional methods. The sensor (Fig. 8.19a) consists of a light-emitting diode (LED) and a hemispherical spirit level mounted on a pn-junction photodiode array. A shadow of the bubble in the liquid is projected onto the surface of the photodiode array. When the sensor is kept horizontally, the shadow on the sensor is circular as shown in Fig. 8.19b, and the area of the shadow on each photodiode of the array is the same. However, when the sensor is inclined, the shadow becomes slightly elliptic as shown in Fig. 8.19c, implying that the output currents from the diodes are no longer equal. In a practical sensor, the diameter of the LED is 10 mm

[6]The Fredericks Company. PO Box 67, Huntingdon Valley, PA 19006.

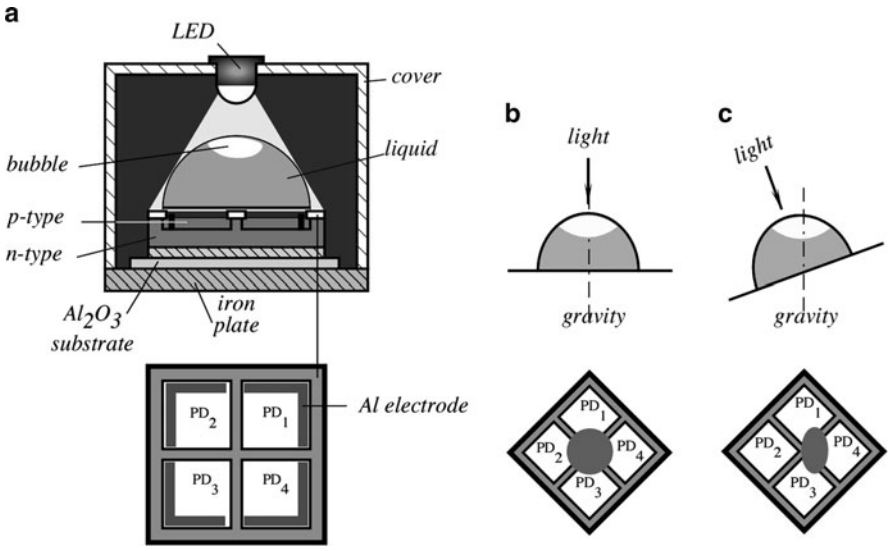

Fig. 8.19 Optoelectronic inclination sensor design (**a**); shadow at a horizontal position (**b**); shadow at the inclined position (**c**)

and the distance between the LED and the level is 50 mm, the diameters of the hemispherical glass and the bubble are 17 and 9 mm, respectively. The outputs of the diodes are converted into a digital form and calibrated at various tilt angles. The calibration data are compiled into look-up tables that are processed by a computing device. By positioning the sensor at the cross point of the lines drawn longitudinally and latitudinally at an interval on the slanting surface of an object, x and y components of the tilt angle can be obtained and the shape of the object is reconstructed by a computer.

References

1. Articolo GA (1989) Shock impulse response of a force balance servo-accelerometer. In: Sensors Expo West proceedings, 1989. © Helmers Publishing, Inc.
2. (1991) Sensor signal conditioning: an IC designer's perspective. Sensors, Nov. 23–30
3. Allen H, Terry S, De Bruin D (1989) Accelerometer system with self-testable features. Sens Actuators 20:153–161
4. Suminto JT (1991) A simple, high performance piezoresistive accelerometer. In: Transducers'91. 1991 international conference on solid-state sensors and actuators. Digest of Technical Papers. pp 104–107, ©IEEE
5. Haritsuka R, van Duyn DS, Otaredian T, de Vries P (1991) A novel accelerometer based on a silicon thermopile. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers. pp 420–423, ©IEEE
6. Fox CHJ, Hardie DSW (1984) Vibratory gyroscopic sensors. In: Symposium gyro technology, DGON

7. Boxenhom BB, Dew B, Greiff P (1989) The micromechanical inertial guidance system and its applications. In: 14th biennial guidance test symposium, 6588th Test Group, Holloman AFB, New Mexico, 3–5 Oct 1989
8. Varnham MP, Hodgins D, Norris TS, Thomas HD (1991) Vibrating planar gyro. US Patent 5,226,321
9. Udd E (1991) Fiber optic sensors based on the Sagnac interferometer and passive ring resonator. In: Udd E (ed) Fiber optic sensors. Wiley, New York, pp 233–269
10. Ezekiel S, Arditty HJ (eds) (1982) Fiber-optic rotation sensors. Springer series in optical sciences, vol 32. Springer, New York
11. Fredericks RJ, Ulrich R (1984) Phase error bounds of fiber gyro with imperfect polarizer/depolarizer. Electron Lett 29:330
12. Bailleul G (1991) Vibracoax piezoelectric sensors for road traffic analysis. In: Sensor Expo proceedings, 1991. ©Helmers Publishing, Inc.
13. Radice PF (1991) Piezoelectric sensors and smart highways. In: Sensors Expo proceedings, 1991. Helmers Publishing, Inc.
14. Ebisawa M, Takeshi N, Tooru S (2007) Coaxial piezoelectric cable polarizer, polarizing method, defect detector, and defect detecting method. US Patent 7,199,508, 3 Apr 2007
15. Piezo film sensors technical manual. Measurement Specialties, Inc. http://www.msiusa.com April 1999
16. Kato H, Kojima M, Gattoh M, Okumura Y, Morinaga S (1991) Photoelectric inclination sensor and its application to the measurement of the shapes of 3-D objects. IEEE Trans Instrum Meas 40(6):1021–1026

Chapter 9
Force, Strain, and Tactile Sensors

Engineering is art of converting science into useful things

While the kinematics studies positions of objects and their motions, the dynamics answers the question – what causes the motion? Classical mechanics deals with moving objects whose velocities are substantially smaller than the speed of light. Moving particles, such as photons, atoms, and electrons, are the subjects of quantum mechanics and the theory of relativity. A typical problem of classical mechanics is the question: "What is motion of an object that initially had a given mass, charge, dipole moment, position, etc. and was subjected to external objects having known mass, charge, velocity, etc.?" That is, classical mechanics deals with interactions of macroobjects. In a general form, this problem was solved by Sir Isaac Newton (1642–1727) who was born in the year when Galileo died. He brilliantly developed ideas of Galileo and other great mechanics. Newton stated his first law as: "*Every body persists in its state of rest or of uniform motion in a straight line unless it is compelled to change that state by forces impressed on it*." Sometimes, this is called a law of inertia. Another way to state the first law is to say that: "*If no net force acts on a body, its acceleration* $\boldsymbol{a}$ *is zero*."

When force is applied to a free body (not anchored to another body), it gives the body an acceleration in a direction of the force. Thus, we can define force as a vector value. Newton had found that acceleration is proportional to the acting force $\mathbf{F}$ and inversely proportional to the property of a body called the mass m, which is a scalar value:

$$a = \frac{\mathbf{F}}{m}. \tag{9.1}$$

This equation is known as Newton's second law – the name given by the great Swiss mathematician and physicist Leonhard Euler in 1752, 65 years after the publication of Newton's Principia [1]. The first law is contained in the second law as a special case: when net acting force $\mathbf{F} = 0$, acceleration $\mathbf{a} = 0$.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_9, © Springer Science+Business Media, LLC 2010

Newton's second law allows us to establish the mechanical units. In SI terms, mass (kg), length (m), and time (s) are the base units (see Table 1.7). Force and acceleration are derivative units. The force unit is the force that will accelerate 1 kg mass to acceleration 1 m/s^2. This unit is called a Newton.

In the British and U.S. Customary Systems of units, however, force (lb), length (ft), and time (s) are selected as the base units. The mass unit is defined as the mass that is accelerated at 1 ft/s^2 when it is subjected to force of 1 pound. The British unit of mass is slug . Hence, the mechanical units are as shown in Table 9.1.

Newton's third law establishes a principle of a mutual interaction between two bodies: "*To every action there is always opposed an equal reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts*."

In engineering measurements, it is often necessary to know the density of a medium, which is amount of matter per unit volume. Density is defined through mass m and volume V as

$$\rho = \frac{m}{V}. \tag{9.2}$$

The unit of density is kg/m^3 or lb/ft^3 (British system). Densities of some materials are given in Table A.12.

The SI unit of force is one of the fundamental quantities of physics. The measurement of force is required in mechanical and civil engineering, for weighing objects, designing prosthesis, etc. Whenever pressure is measured, it requires the measurement of force. It could be said that force is measured when dealing with solids, while pressure – when dealing with fluids (i.e., liquids or gases). That is, force is considered when action is applied to a spot, and pressure is measured when force is distributed over a relatively large area.

Force sensors can be divided into two classes: quantitative and qualitative. A quantitative sensor actually measures the force and represents its value in terms of an electrical signal. Examples of these sensors are strain gauge s and load cells. The qualitative sensors are the threshold devices that are not concerned with a good fidelity of representation of the force value. Their function is merely to indicate whether a sufficiently strong force is applied or not. That is, the output signal indicates when the force magnitude exceeds a predetermined threshold level. An example of these detectors is a computer keyboard where a key makes a contact only when it is pressed sufficiently hard. The qualitative force sensors are frequently used for detection of motion and position, as described in Chaps. 7 and 8. A pressure sensitive floor mat and a piezoelectric cable are examples of the qualitative force sensors.

Table 9.1 Mechanical units (bold face indicates base units)

System of units	Force	Mass	Acceleration
SI	Newton (N)	**Kilogram** (kg)	m/s^2
British	**Pound** (lb)	Slug	ft/s^2

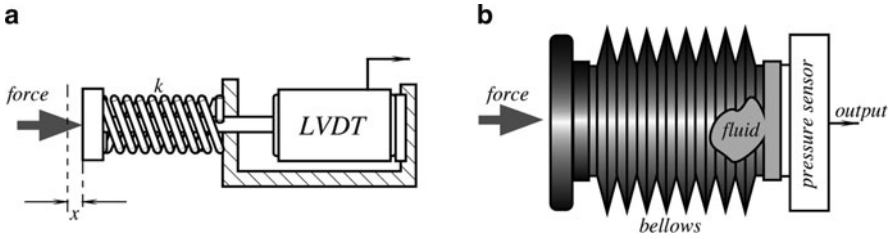

Fig. 9.1 Spring-loaded force sensor with LVDT (**a**). Force sensor incorporating a pressure sensor (**b**)

The various methods of sensing force can be categorized as follows [2]:

1. By balancing the unknown force against the gravitational force of a standard mass
2. By measuring the acceleration of a known mass to which the force is applied
3. By balancing the force against an electromagnetically developed force
4. By converting the force to a fluid pressure and measuring that pressure
5. By measuring the strain produced in an elastic member by the unknown force

In the modern sensors, the most commonly used method is 5, while 3 and 4 are used occasionally.

In most sensors, force is not directly converted into an electric signal. Some intermediate steps are usually required. Thus, many force sensors are the complex sensors. For instance, a force sensor can be fabricated by combining a force-to-displacement transducer and a position (displacement) sensor. The former may be a simple coil spring, whose compression displacement x can be defined through the spring coefficient k and compressing force F as

$$x = kF. \tag{9.3}$$

The sensor shown in Fig. 9.1a is comprised of a spring and LVDT displacement sensor (Sect. 7.4). Within the linear range of the spring, the LVDT sensor produces voltage, which is proportional to the applied force. A similar sensor can be constructed with other types of springs and pressure sensors, such as the one shown in Fig. 9.1b. The pressure sensor is combined with a fluid-filled bellows, which is subjected to force. The fluid-filled bellows functions as a force-to-pressure converter (transducer) by distributing a localized force at its input over the sensing membrane of a pressure sensor.

9.1 Strain Gauges

Strain is deformation of a physical body under the action of applied forces. A strain gauge is a resistive elastic sensor whose resistance is function of the applied strain (unit deformation). Since all materials resist to deformation, some force must be

applied to cause deformation. Hence, resistance can be related to applied force. That relationship is generally called the piezoresistive effect (see Sect. 3.5.3) and is expressed through the gauge factor S_e of the conductor (3.63):

$$\frac{dR}{R} = S_e e. \tag{9.4}$$

For many materials $S_e \approx 2$ with the exception of platinum for which $S_e \approx 6$ [3]. For small variations in resistance not exceeding 2% (which is usually the case), the resistance of the metallic wire can be approximated by a linear equation:

$$R = R_o(1+x), \tag{9.5}$$

where R_o is the resistance with no stress applied, and $x = S_e e$. For the semiconductive materials, the relationship depends on the doping concentration (Fig. 18.2a). Resistance decreases with compression and increases with tension. Characteristics of some resistance strain gauges are given in Table 9.2.

A wire strain gauge is composed of a resistor bonded with an elastic carrier (backing). The backing, in turn, is applied to the object where stress or force should be measured. Obviously, that strain from the object must be reliably coupled to the gauge wire, while the wire must be electrically isolated from the object. The coefficient of thermal expansion of the backing should be matched to that of the wire. Many metals can be used to fabricate strain gauges. The most common materials are alloys constantan, nichrome, advance, and karma. Typical resistances vary from 100 to several thousand ohms. To possess good sensitivity, the sensor should have long longitudinal and short transverse segments (Fig. 9.2), so that transverse sensitivity is no more than a couple of percent of the longitudinal. The gauges may be arranged in many ways to measure strains in different axes. Typically, they are connected into Wheatstone bridge circuits (Sect. 5.10). It should be noted, that semiconductive strain gauges are quite sensitive to temperature variations. Therefore, interface circuits or the gauges must contain temperature compensating networks.

Table 9.2 Characteristics of some resistance strain gauges (after [4])

Material	Gauge factor (S_e)	Resistance (Ω)	TCR (°C−1 × 10−6)	Notes
57%Cu–43%Ni	2.0	100	10.8	S_e is constant over wide range of strain. For use under 260°C.
Platinum alloys	4.0–6.0	50	2,160	For high temperature use
Silicon	−100 to +150	200	90,000	High sensitivity, good for large strain measurements

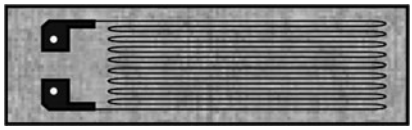

Fig. 9.2 Wire strain gauge bonded to elastic backing

9.2 Tactile Sensors

The tactile sensors loosely can be subdivided into three subgroups:

Touch Sensors. These sensors detect and measure contact forces at defined points. A touch sensor typically is a threshold device or a binary sensor, namely – touch or no touch.

Spatial Sensors. These sensors detect and measure the spatial distribution of forces perpendicular to a predetermined sensory area, and the subsequent interpretation of the spatial information. A spatial-sensing array can be considered to be a coordinated group of touch sensors.

Slip Sensors. These sensors detect and measure the movement of an object relative to the sensor. This can be achieved either by a specially designed slip sensor or by the interpretation of the data from a touch sensor or a spatial array.

In general, the tactile sensors are a special class of force or pressure transducers that are characterized by small thickness. This makes the sensors useful in the applications where force or pressure can be developed between two surfaces being in close proximity to one another. Examples include robotics where tactile sensors can be positioned on the "fingertips" of a mechanical actuator to provide a feedback upon developing a contact with an object – very much like tactile sensors work in human skin. They can be used to fabricate "touch screen" displays, keyboards, and other devices where a physical contact has to be sensed. A very broad area of applications is in the biomedical field where tactile sensors can be used in dentistry for the crown or bridge occlusion investigation, in studies of forces developed by a human foot during locomotion. They can be installed in artificial knees for the balancing of the prosthesis operation, etc. In mechanical and civil engineering, the sensors can be used to study forces developed by fastening devices.

Some touch sensors do not rely on reaction to a force. A touch by a finger may be detected by monitoring a contact area between the finger and the panel. An example is a touch screen on a mobile telephone.

Requirements to tactile sensors are based on investigation of human sensing and the analysis of grasping and manipulation. An example of the desirable characteristics of a touch or tactile sensor suitable for the majority of industrial applications is as follows:

1. A touch sensor should ideally be a single-point contact, though the sensory area can be any size. In practice, an area of 1–2 mm^2 is considered a satisfactory.
2. The sensitivity of the touch sensor is dependent on a number of variables determined by the sensor's basic physical characteristics. In addition, the sensitivity

depends on the application, in particular, any physical barrier between the sensor and the object. A sensitivity within the range 0.4–10 N, together with an allowance for accidental mechanical overload, is considered satisfactory for most industrial applications.

3. A minimum sensor bandwidth of 100 Hz.
4. The sensor characteristics must be stable and repeatable with low hysteresis. A linear response is not absolutely necessary as information processing techniques can be used to compensate for any moderate nonlinearities.

If a tactile array is being considered, the majority of applications can be undertaken by an array of 10–20 sensors square, with a spatial resolution of 1–2 mm. In robotics and in design of prosthesis, a grasping force at a "finger" tip should be measured. Thus, these tactile sensors can be integrated into the "skin" to respond in real time, the magnitude, location, and direction of the forces at the contact point.

9.2.1 Switch Sensors

Several methods can be used to fabricate tactile sensors. Some of them require a formation of a thin layer of a material, which is responsive to strain. A simple tactile sensor producing an "on–off" output can be formed with two leaves of foil and a spacer (Fig. 9.3). The spacer has round (or any other suitable shape) holes. One leaf is grounded and the other is connected to a pull-up resistor. A multiplexer can be used if more than one sensing area is required. When an external force is applied to the upper conductor over the hole in the spacer, the top leaf flexes and upon reaching the lower conductor, makes an electric contact, grounding the pull-up resistor. The output signal becomes zero indicating the applied force. The upper and lower conducting leaves can be fabricated by a silk-screen printing of conductive ink on the backing material, like Mylar® or polypropylene. Multiple sensing spots can be formed by printing rows and columns of a conductive ink. Touching of a

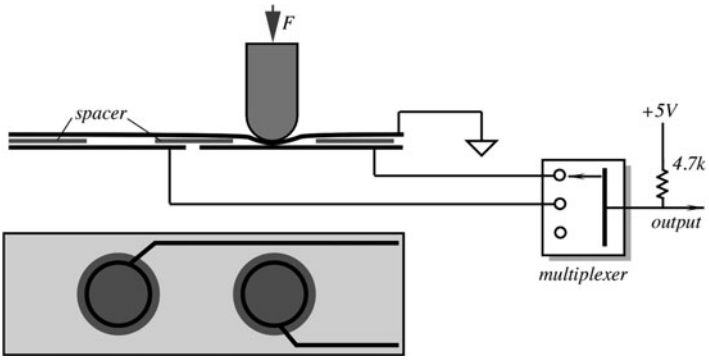

Fig. 9.3 Membrane switch as a tactile sensor

particular area on a sensor will cause the corresponding row and column to join thus indicating force at a particular location.

9.2.2 *Piezoelectric Sensors*

Good tactile sensors can be designed with piezoelectric films , such as polyvinylidene fluoride (PVDF) used in active or passive modes. An active ultrasonic coupling touch sensor with the piezoelectric films is illustrated in Fig. 9.4 where three films are laminated together (the sensors also has additional protective layers, which are not shown in the figure). The upper and the bottom films are PVDF, while the center film is for the acoustic coupling between the other two. The softness of the center film determines sensitivity and the operating range of the sensor. The bottom piezoelectric film is driven by an AC voltage from an oscillator. This excitation signal results in mechanical contractions of the film that are coupled to the compression film and, in turn, to the upper piezoelectric film, which acts as a receiver. Since piezoelectricity is a reversible phenomenon, the upper film produces alternating voltage upon being subjected to mechanical vibrations from the compression film. These oscillations are amplified and fed into a synchronous demodulator. The demodulator is sensitive to both the amplitude and the phase of the received signal. When compressing force F is applied to the upper film, mechanical coupling between the three-layer assembly changes. This affects the amplitude and the phase of the received signal. These changes are recognized by the demodulator and appear at its output as a variable voltage.

Within certain limits, the output signal linearly depends on the force. If the 25-μm PVDF films are laminated with a 40-μm silicone rubber compression film, the thickness of an entire assembly (including protective layers) does not exceed 200 μm. The PVDF film electrodes may be fabricated with a cell-like pattern on

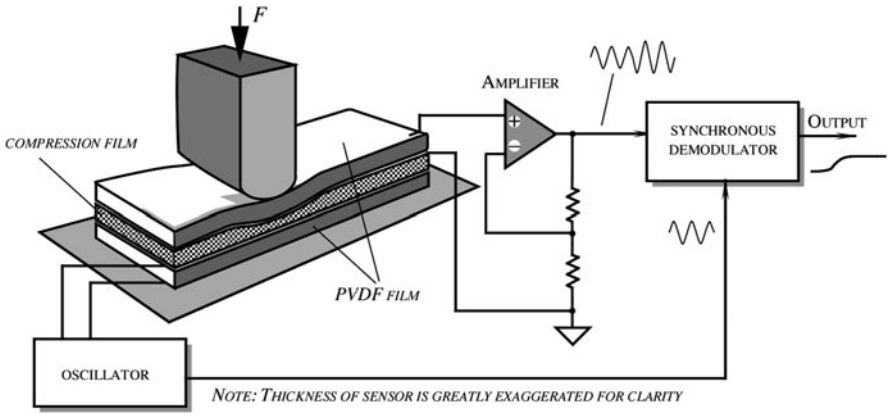

Fig. 9.4 Active piezoelectric tactile sensor

either the transmitting or receiving side. This would allow using an electronic multiplexing of the cells to achieve spatial recognition of the applied stimuli. The sensor also can be used to measure small displacement s. Its accuracy is better than ± 2 μm over a few millimeter ranges. The advantages of this sensor are in its simplicity and a DC response, that is, in the ability to recognize static forces.

A piezoelectric tactile sensor can be fabricated with the PVDF film strips imbedded into a rubber skin (Fig. 9.5a). This sensor is passive, that its, its output signal is generated by the piezoelectric film without the need for an excitation signal. As a result, it produces a response proportional to the rate of stress, rather than to the stress magnitude. A design of this sensor is geared to robotic applications where it is desirable to sense sliding motions causing fast vibrations. The piezoelectric sensor is directly interfaced with a rubber skin, thus the electric signal produced by the strips reflect movements of the elastic rubber, which results from the friction forces.

The sensor is built on a rigid structure (a robot's "finger"), which has a foamy compliant underlayer (1 mm thick), around which a silicon rubber "skin" is wrapped. It is also possible to use a fluid underlayer for a better smooth surface tracking. Because the sensing strips are located at some depth beneath the skin surface, and because the piezoelectric film responds differently in different directions, a signal magnitude is not the same for movements in any direction. The sensor responds with a bipolar signal (Fig. 9.5b) to surface discontinuity or bumps as small as 50 μm high.

Here are few more examples of sensors that use the PVDF and copolymer films [6]. Many tactile sensors are just sensitive conventional switches. However, the reliability of conventional contact switches is reduced due to contaminants like moisture and dust that foul the contact points. Piezoelectric film offers exceptional reliability as it is a monolithic structure, not susceptible to these and other conventional switch failure modes. One of the most challenging of all switch applications is found in pinball machines. A pinball machine manufacturer uses a piezo film switch as a replacement for the momentary rollover type switch. The switch is constructed from a laminated piezoelectric film on a spring steel beam, mounted as a cantilever to the end of a circuit board (Fig. 9.6a). The "digital" piezoelectric film switch is connected to a simple MOSFET circuit that consumes no power during the

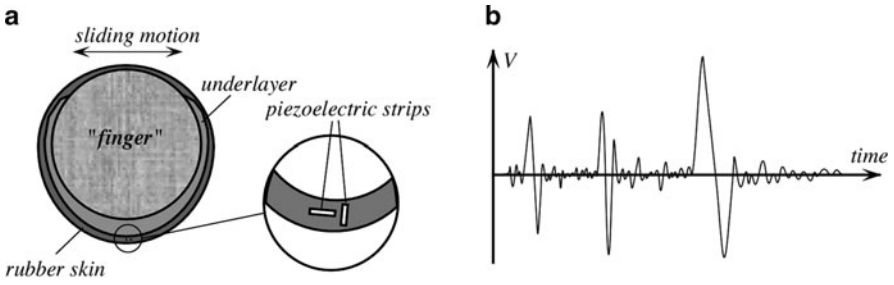

Fig. 9.5 Tactile sensor with a piezoelectric film for detecting sliding forces cross-sectional view (**a**); typical response (**b**) (adapted from [5])

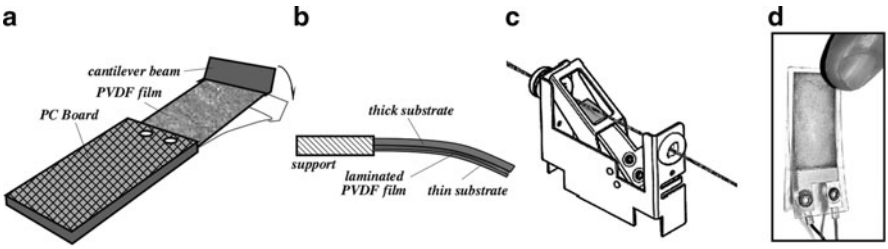

Fig. 9.6 PVDF film switch for a pinball machine (**a**), beam switch (**b**), thread break sensor (**c**) (from [6]), laminated tactile sensor (**d**)

normally open state. In response to a direct contact force, the film beam flexes, generates electric charge and momentarily triggers the MOSFET. This provides a momentary "high" state of the switch. The sensor does not exhibit the corrosion, pitting, or bounce that are normally associated with contact switches. It can survive in excess of ten million cycles without failure. The simplicity of the design makes it effective in applications that include: counter switches for assembly lines and shaft rotation, switches for automated processes, impact detection for machine dispensed products, etc. The cantilever beam that carries the PVDF film can be modified to adjust switch sensitivity from high to low impact forces. Figure 9.6b shows the construction of the beam-type switch. The PVDF film element is laminated to a thicker substrate on one side, and has a much thinner laminated substrate on the other. This moves the neutral axis of the structure out of the piezoelectric film element, resulting in a fully tensile strain in the film when deflected downward, and a fully compressive strain when deflected in the opposite direction. Beam switches are used in shaft rotation counters in natural gas meters and as gear tooth counters in electric utility metering. The beam switch does not require an external power source, so the gas meter is safe from spark hazard. Other examples of applications for the beam switch include a baseball target that detects ball impact, a basketball game where a hoop-mounted piezoelectric film sensor counts good baskets, switches inside of an interactive soft doll to detect a kiss to the cheek or a tickle (the sensor is sewn into the fabric of the doll), coin sensors for vending and slot machines, and as digital potentiometer for high reliability.

The popularity of electronics for musical instruments presents a special problem in drums and pianos. The very high dynamic range and frequency response requirements for the drum triggers and piano keyboards are met by piezoelectric film impact elements. Laminates of piezo film are incorporated in the foot pedal switches for bass drums, and triggers for snares and tom-toms. Piezoelectric film impact switches are force sensitive, faithfully duplicating the effort of the drummer or pianist. In electronic pianos, the piezoelectric film switches respond with a dynamic range and time constant that is remarkably similar to a piano keystrokes.

Textile plants require the continuous monitoring of often thousands of lines of thread for breakage. An undetected break event can require that a large volume of material be discarded, as the labor costs to recover the material exceed the

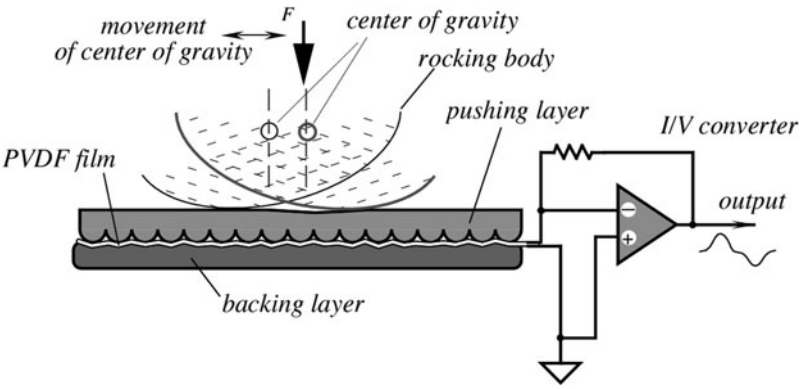

Fig. 9.7 Piezoelectric film respiration sensor

manufacturing cost. Drop switches, where switch contact closure occurs when the thread breaks, are very unreliable. Lint fouls the contact points, resulting in no output signal. A piezoelectric film vibration sensor, mounted to a thin steel beam, monitors the acoustic signal caused by the abrasion of the thread running across the beam, analogous to a violin string (Fig. 9.6c). The absence of the vibration instantly triggers the machinery to stop.

Figure 9.7 shows a PVDF film tactile sensor for detecting breathing rate of a sleeping child, where minute movements of a body resulted from respiration had to be monitored in order to detect cessation of breathing [7]. The sensor was placed under the mattress in a crib. A body of a normally breathing child slightly shifts with each inhale and exhale due to a moving diaphragm. This results in a displacement of the body's center of gravity that is detected by the PVDF film sensor. The sensor consists of three layers where the PVDF film is positioned between a backing material (for instance, silicone rubber) and a pushing layer.

The pushing layer is fabricated of a plastic film (for instance, Mylar®) whose side facing the PVDF film is preformed to have a corrugated surface. Under the variable force, the PVDF film is variably stressed by the groves of the pusher. In response, the film generates an electric current. The current flows through a current-to-voltage (*I*/*V*) converter that produces a variable output voltage. The amplitude of that voltage within certain limits is proportional to the applied gravitational force.

9.2.3 Piezoresistive Sensors

Another type of a tactile sensor is a piezoresistive sensor. It can be fabricated by using materials whose electrical resistance is function of strain. The sensor incorporates a force-sensitive resistor (FSR) whose resistance varies with applied

pressure [8]. Such materials are conductive elastomers or pressure-sensitive inks. A conductive elastomer is fabricated of silicone rubber, polyurethane, and other compounds that are impregnated with conductive particles or fibers. For instance, conductive rubber can be fabricated by using carbon powder as an impregnating material. Operating principles of elastomeric tactile sensors are based either on varying the contact area when the elastomer is squeezed between two conductive plates (Fig. 9.8a) or in changing the thickness. When the external force varies, the contact area at the interface between the pusher and the elastomer changes, resulting in a reduction of electrical resistance.

At a certain pressure, the contact area reaches its maximum and the transfer function (Fig. 9.8b) goes to saturation. For a resistive polymer Velostat™ (from 3M) having thickness 70 μm and a specific resistance of 11 kΩ/cm^2, resistance for pressures over 16 kPa can me approximated by equation

$$R = \frac{51.93}{p^{1.47}} + 19. \tag{9.6}$$

It should be noted, however, that the resistance may noticeably drift when the polymer is subjected to prolonged pressure. Thus, the FSR sensors would be much more useful for the qualitative rather than quantitative measurements.

A much thinner FSR tactile sensor can be fabricated with a semiconductive polymer whose resistance varies with pressure. A design of the sensor resembles a membrane switch (Fig. 9.9) [9]. Compared with a strain gauge, the FSR has a much wider dynamic range: typically three decades of resistance change over a 0–3 kg force range and much lower accuracy (typically ±10%). However, in many applications, where an accurate force measurement is not required, a very low cost of the sensor makes it an attractive alternative. A typical thickness of a FSR polymer sensor is on the range of 0.25 mm (0.010″) but much thinner sheets are also available.

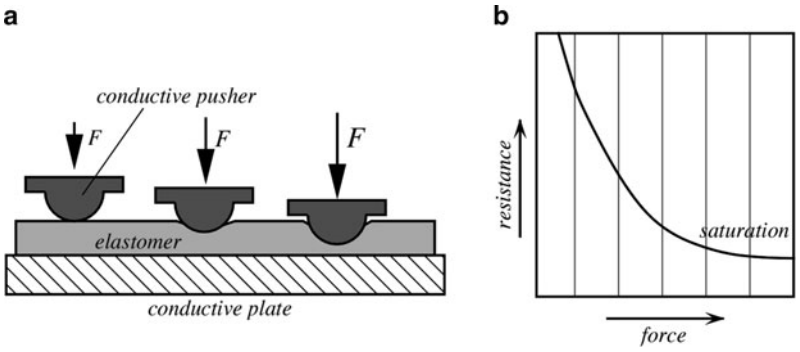

Fig. 9.8 FSR tactile sensor through-thickness application with an elastomer (**a**); transfer function (**b**)

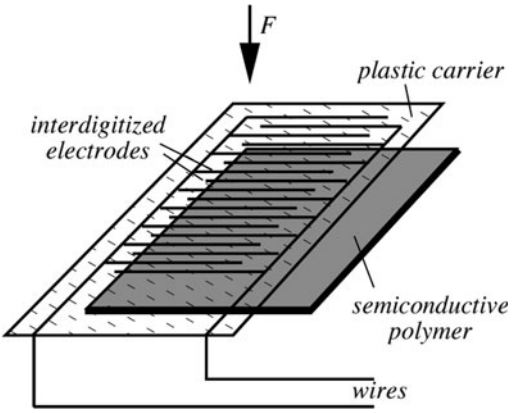

Fig. 9.9 Tactile sensor with a polymer FSR

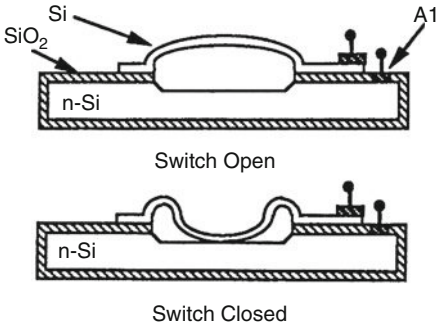

Fig. 9.10 Micromachined silicon threshold switch with trapped gas (from [10])

9.2.4 MEMS Sensors

Miniature tactile sensors are especially in high demand in robotics, where good spatial resolution, high sensitivity, and a wide dynamic range are required. A plastic deformation in silicon can be used for the fabrication of a threshold tactile sensor with a mechanical hysteresis. In one design [10], the expansion of trapped gas in a sealed cavity formed by wafer bonding is used to plastically deform a thin silicon membrane bonded over the cavity, creating a spherically shaped cap. The structure shown in Fig. 9.10 is fabricated by a micromachining technology of a silicon wafer. At normal room temperature and above a critical force, the upper electrode will buckle downward, making contact with the lower electrode.

Experiments have shown that the switch has hysteresis of about 2 psi of pressure with a closing action near 13 psi. The closing resistance of the switch is on the order of 10 kΩ, which for the micropower circuits is usually low enough.

In another design, a vacuum, instead of pressurized gas, is used in a microcavity. This sensor shown in Fig. 9.11 [11] has a silicon vacuum configuration, with a cold

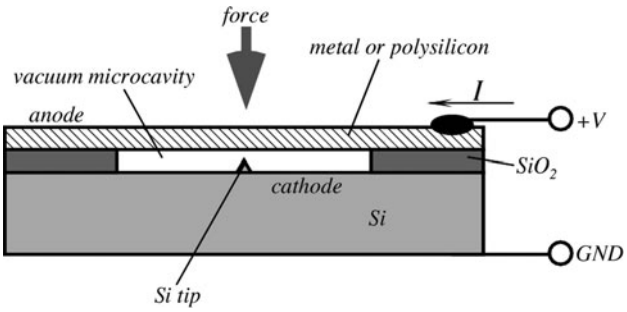

Fig. 9.11 Schematic of a vacuum diode force sensor (adapted from [11])

field emission cathode and a movable diaphragm anode. The cathode is a sharp silicon tip. When a positive potential difference is applied between the tip and the anode, an electric field is generated, which allows electrons to tunnel from inside the cathode to the vacuum, if the field exceeds 5×10^7 V/cm [12]. The field strength at the tip and quantity of electrons emitted (emission current) are controlled by the anode potential. When an external force is applied, the anode deflects downward, thus changing the field and the emission current.

The emission current can be expressed through the anode voltage V as

$$I = V^2 a \exp\left(-\frac{b}{\beta V}\right), \tag{9.7}$$

where a and b are constants, and β is the tip geometry factor, which depends on the distance between the anode and the cathode. To achieve a better sensitivity, the tip is fabricated with a radius of curvature of about 0.02 μm.

9.2.5 Capacitive Touch Sensors

A capacitive touch sensor is based on fundamental equations for the parallel-plate and coaxial capacitors (Sect. 3.2.1). A capacitive touch sensor relies on the applied force that either changes the distance between the plates or the variable surface area of the capacitor. In such a sensor, two conductive plates are separated by a dielectric medium, which is also used as the elastomer to give the sensor its force-to-capacitance characteristics (Fig. 9.12a).

To maximize the change in capacitance as force is applied, it is preferable to use a high permittivity dielectric in a coaxial capacitor design (Fig. 9.12b). The use of a highly dielectric polymer such as PVDF maximizes the change capacitance. From an application viewpoint, the coaxial design is better as its capacitance will give a greater increase for an applied force than the parallel-plate design; however, the sensor is more complex and larger.

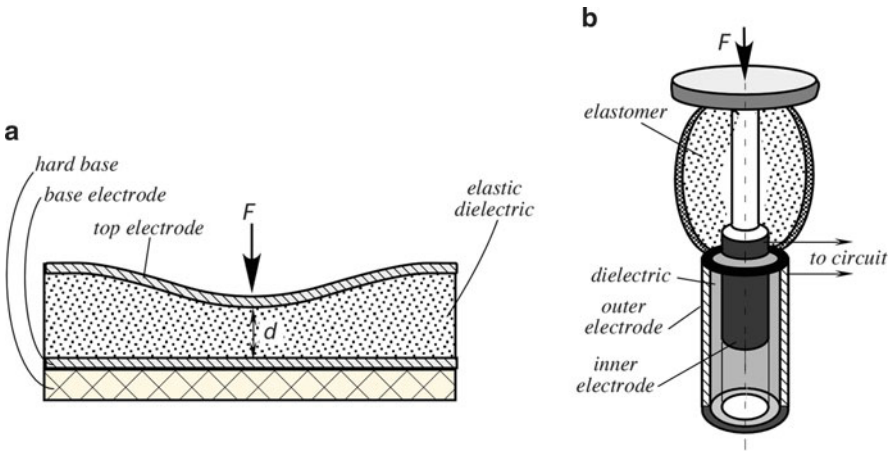

Fig. 9.12 Capacitive tactile sensors. Parallel-plate sensor (**a**) and coaxial sensor (**b**)

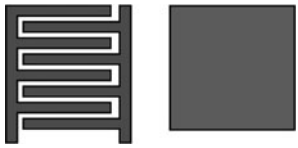

Fig. 9.13 Interdigitized and single electrodes for a touch screen

To measure the change in capacitance, a number of techniques can be employed. The most popular is based on the use of a current source with a resistor and measure the time delay caused by a variable capacitance. A second approach is to use the sensor as part of an oscillator with an LC or RC circuit, and measure the frequency response. Significant problem with capacitive sensors can be caused if they are in close proximity with the metal structures, this leads to stray capacitance. The effect can be minimized by a good circuit layout and mechanical design. It is possible to fabricate a parallel-plate capacitor on a single silicon slice; this can give a very compact sensing device. When a reference capacitor is employed, a practical circuit that is shown in Fig. 5.32 may come in handy.

The capacitive sensors are popular in touch screen panels that typically are made of glass or a clear polymer coated with a transparent conductor such as indium tin oxide (ITO) that combines electrical conductivity and optical clarity. This type of sensor is basically a capacitor in which the plates are the overlapping areas between the horizontal and vertical axes in a grid pattern. Each plate may be a dual interdigitized or single electrode (Fig. 9.13). Since the human body also conducts electricity, a touch on the surface of the sensor will affect the electric field and create a measurable change in the capacitance of the device. These sensors work on proximity of the conductive medium (finger), and do not have to be directly touched to be triggered. It is a durable technology that is used in a wide range of applications including point-of-sale systems, industrial controls, and public information kiosks.

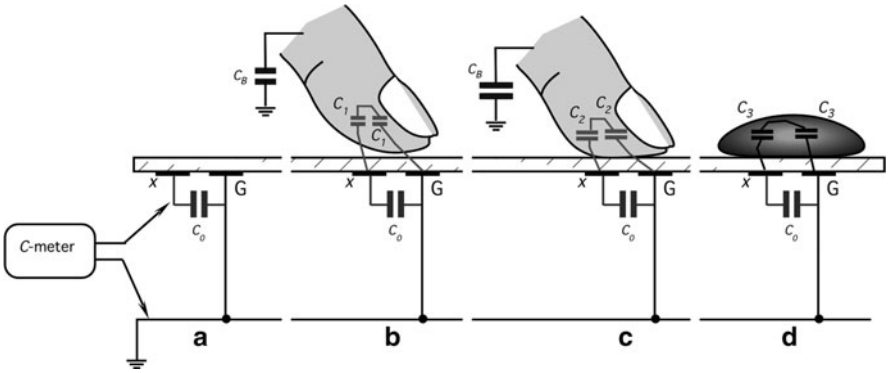

Fig. 9.14 A dual-electrode touch screen. No touch (**a**), light touch (**b**), strong touch (**c**), and a water droplet (**d**)

However, it only responds to finger contact and will not work with a gloved hand or pen stylus unless the stylus is conductive.

Consider two electrodes deposited on a glass screen as shown in Fig. 9.14a. One of the electrodes (G) is grounded and the other is connected to the capacitance meter (*C*-meter). Some small baseline capacitance C_0 exists between the two electrodes and that capacitance is monitored by the *C*-meter.

When a finger comes in proximity of the electrodes (Fig. 9.14b), it develops a capacitive coupling C_1 with each electrode. In response the capacitance, monitor will register a new combined capacitance

$$C_{m1} = C_0 + 0.5C_1, \tag{9.8}$$

which is much larger than C_0. If the finger is pressed harder, because of the fingertip elasticity, the contact area with the touch screen increases and that causes a larger capacitive coupling $C_2 > C_1$ as shown in Fig. 9.14c. This will further increase the combined monitored capacitance and thus can be used as an indication of a harder pressing. Now, let us assume that a droplet of water is deposited on the touch screen above the electrodes as shown in Fig. 9.14d. Being electrically conductive with a dielectric constant between 76 and 80, water forms a strong coupling C_3 with the electrodes, which is comparable with that of a finger and, as a result, the touch screen will indicate a false touch. Sensitivity to water droplets is a disadvantage of a dual-electrode touch screen where one electrode is grounded.

To resolve a sensitivity to water droplets, an improvement of a capacitive touch screen that contains a single-electrode pattern was proposed [13]. No electrode in that screen is grounded. Under the no-touch condition, only a small capacitance Cg exists between the electrode and ground (earth) as indicated in Fig. 9.15a and it is monitored by the *C*-meter. A human body naturally forms a strong capacitive coupling C_B to the surrounding objects. This capacitance is several orders of magnitude larger than C_0. Hence, a human body may be considered a "ground." When a finger comes in the vicinity of the electrode (Fig. 9.15b), a capacitance C_1 is

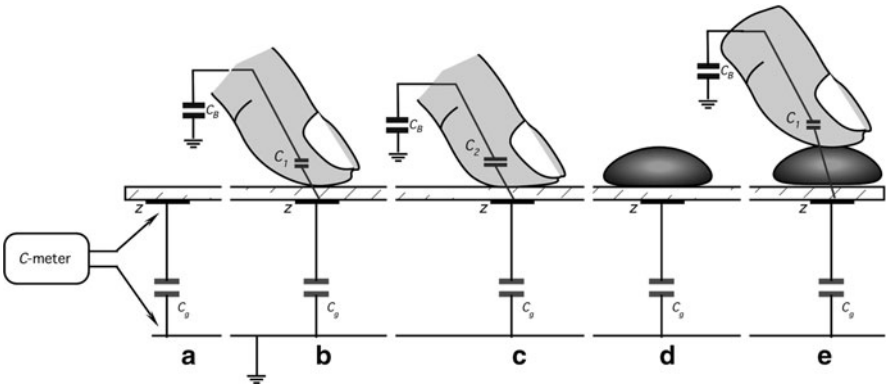

Fig. 9.15 A single-electrode touch screen. No touch (**a**), light touch (**b**), strong touch (**c**), a water droplet (**d**), and touching through a water droplet (**e**)

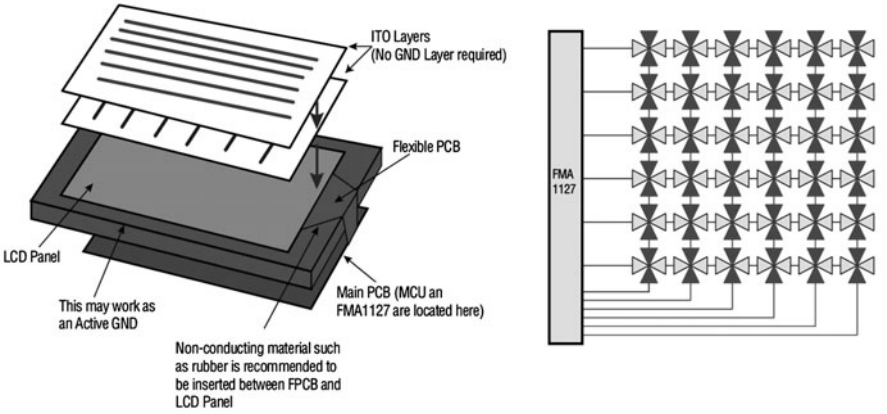

Fig. 9.16 Nongrounded touch screen for LCD panel (Courtesy of Fujitsu Microelectronics America, Inc.)

formed between the finger tip and the electrode. This capacitance is electrically connected in parallel to the baseline capacitance C_0, causing the C-meter to respond. Like in a two-electrode screen, a stronger pressing will create a larger capacitance as shown in Fig. 9.15c. However, when a water droplet is deposited on the screen, it will cause no detection as the water droplet is not couplet to ground as shown in Fig. 9.15d. It is interesting to note that as shown in Fig. 9.15e touching the water droplet will form a capacitive coupling to ground and the touch will be correctly detected. Therefore, this electrode arrangement is more robust to the environmental conditions. Arranging the electrode pattern in rows and columns and processing signals by the appropriate circuit can make a reliable spatial touch recognition by an appropriate electronic circuit such as the Fujitsu controller FMA1127 shown in Fig. 9.16.

A similar approach can be used to form proximity detectors on surfaces of many shapes. For example, a proximity sensor can be formed on a door knob for the security purposes. It will respond not only to touching but even to approaching the door knob surface by as far as 5 cm.

9.2.6 Acoustic Touch Sensors

ReverSys®[1] acoustic touch screen is based on the recognition of sound waves propagating in an object when the user touches it [14]. A touch of an object produces a pattern of sound waves propagating through the material. This pattern creates an acoustic signature that is unique to the location of the impact. This property is called Time Reversal Acoustics, which can be used to precisely identify location of the source of radiated waves. An acoustic sensor picks up the vibration in the material and passes them to a microcontroller that captures the audio vibrations within an object, generates, and stores the acoustic signatures. This is done during the training of the sensor. In use, when touching the object at the same spot that was previously touched in training, the detected acoustic pattern is compared with the database of signatures and matches a response to the stored pattern.

Another acoustic touch sensor are based on the surface acoustic waves (SAW) technology that uses ultrasonic waves passing over the touch screen panel. When the panel is touched, a portion of the wave is absorbed. This change in the ultrasonic waves registers the position of the touch event and sends this information to the controller for processing. Surface wave touch screen panels can be damaged by outside elements. Contaminants on the surface such as water or oil droplets can also interfere with the functionality of the touch screen.

9.2.7 Optical Sensors

Conventional optical-touch systems use an array of infrared (IR) light-emitting diodes (LEDs) on two adjacent bezel edges of a display, with photo detectors placed on the two opposite bezel edges to analyze the system and determine a touch event (Fig. 9.17). The LED and photo detectors pairs create a grid of light beams across the display. An object (such as a finger or pen) that touches the screen changes the reflection due to a difference between refractive properties of air and a finger. This results in a measured decrease in light intensity at the corresponding photo detector. The measured photo detector outputs can be used to locate a touch-point coordinate.

[1] www.sensitiveobject.fr.

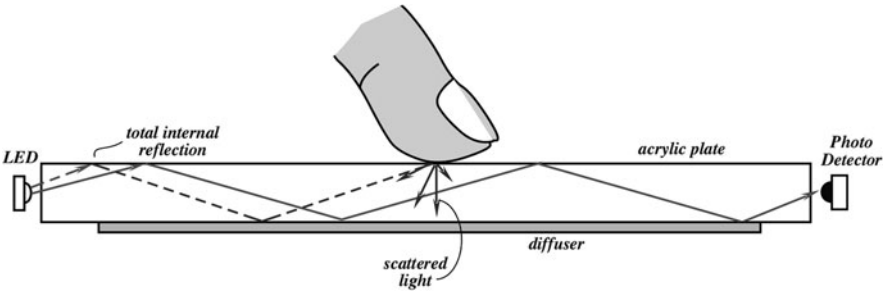

Fig. 9.17 Concept of optical touch screen

Widespread adoption of infrared touch screens has been hampered by two factors: the relatively high cost of the technology compared to competing capacitive technologies and somewhat reduced performance in bright ambient light. The latter problem is a result of background light increasing the noise floor at the photo detectors, sometimes to such a degree that the touch screen's LED light cannot be detected at all, causing a temporary failure of the touch screen. This is most pronounced in direct sunlight conditions where the sun has a very high energy distribution in the infrared region.

However, certain features of IR touch remain desirable and represent attributes of the ideal touch screen, including the option to eliminate the glass or plastic overlay that most other touch technologies require in front of the display. In many cases, this overlay is coated with an electrically conducting transparent material such as ITO, which reduces the optical quality of the display. This advantage of optical touch screens is very important for many applications that require higher clarity.

9.3 Piezoelectric Force Sensors

While the tactile sensors that use piezoelectric effect as it was described above are not intended for the precision measurement of force, the same effect can be used quite efficiently for the precision measurements in different sensor designs. Piezoelectric effects can be used in both passive and active force sensors. It should be remembered however that a piezoelectric effect is, so to speak, an AC effect. In other words, it can convert a changing force into a changing electrical signal, while a steady-state force produces no electrical response. Yet, force can change some properties of a material that would affect an AC piezoelectric response when a sensor is supplied by an active excitation signal. One example of an active approach is shown in Fig. 9.4. However, for quantitative measurements, this would not be a precision approach. A better design would be based on the effect of an applied force modulating the mechanical resonant of the piezoelectric crystal. A basic idea behind the sensor's operation is that certain cuts of quartz crystal,

when used as resonators in electronic oscillators, shift the resonant frequency upon being mechanically loaded. The equation describing the natural mechanical frequency spectrum of a piezoelectric oscillator is given by [15]

$$f_n = \frac{n}{2l}\sqrt{\frac{c}{\rho}}, \tag{9.9}$$

where n is the harmonic number, l is the resonance-determining dimension (e.g., the thickness of a relatively large thin plate or the length of a thin long rod), c is the effective elastic stiffness constant (e.g., the shear stiffness constant in the thickness direction of a plate or Young's modulus in the case of a thin rod), and ρ is the density of the crystal material.

The frequency shift induced by external force is due to nonlinear effects in the crystal. In the above equation, the stiffness constant c changes slightly with the applied stress. The effect of the stress on the dimension (strain) or the density is negligible. The minimal sensitivity to external force can occur when the squeezed dimension is aligned in certain directions for a given cut. These directions are usually chosen when crystal oscillators are designed, because their mechanical stability is important. However, in the sensor applications, the goal is just the opposite – a sensitivity to force along certain axes should be maximized. For example, the diametric force has been used for a high-performance pressure transducer [16] (Fig. 9.18).

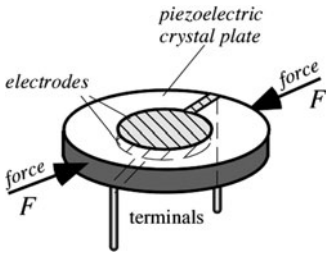

Fig. 9.18 A piezoelectric disk resonator as a diametric force sensor

Another design of a sensor that operates over a relatively narrow range from 0 to 1.5 kg, however, with a good linearity and over 11-bit resolution is shown in Fig. 9.19. quartz To fabricate the sensor, a rectangular plate is cut of the crystal where only one edge is parallel to the x axis, and the face of the plate is cut at the angle of approximately $\Theta = 35^\circ$ with respect to the z axis. This cut is commonly known as AT-cut plate (Fig. 9.19a).

The plate is given surface electrode s for utilizing a piezoelectric effect (see Fig. 3.22), which are connected in a positive feedback of an oscillator (Fig. 9.19b). A quartz crystal oscillates at a fundamental frequency f_o (unloaded), which shifts at loading by [17]

$$\Delta f = F\frac{Kf_o^2 n}{l}, \tag{9.10}$$

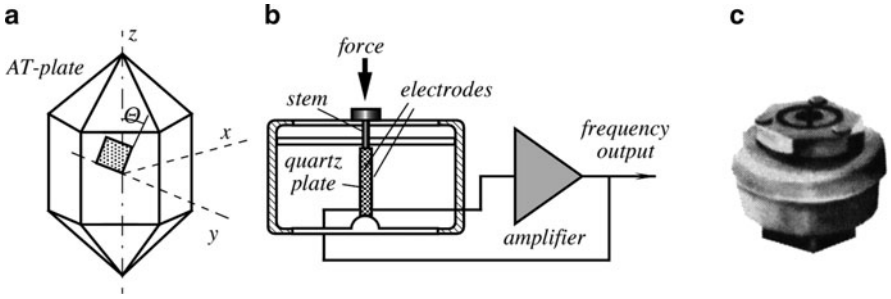

Fig. 9.19 Quartz force sensor AT-cut of a quartz crystal (**a**); structure of the sensor (**b**); outside appearance (**c**) (Courtesy of Quartzcell, Santa Barbara, CA)

where F is the applied force, K is a constant, n is the number of the overtone mode, and l is the size of the crystal. To compensate for frequency variations due to temperature effects, a double crystal can be employed, where one half is used for temperature compensation. Each resonator is connected into its own oscillating circuit and the resulting frequencies are subtracted, thus negating a temperature effect. A commercial force sensor is shown in Fig. 9.14c.

A fundamental problem in all force sensors that use crystal resonators is based on two counter-balancing demands. On one hand, the resonator must have the highest possible quality factor, which means the sensor shall be decoupled from the environment and possibly should operate in vacuum. On the other hand, application of force or pressure requires relatively rigid structure and substantial loading effect on the oscillation crystal, thus reducing its quality factor.

This difficulty may be partly solved by using a more complex sensor structure. For instance, in a photolithographically produced double- and triple-beam structures [15, 18], the so-called "string" concept is employed. The idea is to match dimensions of the oscillating element to the acoustic quarter wavelength ($1/4\lambda$). The total wave reflection occurs at the supporting points through which the external force is coupled and the loading effect on quality factor is significantly reduced.

References

1. Raman VV (1972) The second law of motion and Newton equations. Phys Teacher
2. Doebelin EO (1966) Measurement systems: applications and design. McGraw-Hill, New York
3. Pallás-Areny R, Webster JG (2001) Sensors and signal conditioning, 2nd edn. Wiley, New York
4. Holman JP (1978) Experimental methods for engineers. McGraw-Hill, New York
5. Howe RT (1988) Surface micromachining for microsensors and microactuators. J Vac Sci Technol B 6(6):1809–1813
6. Piezo film sensors technical manual (1999) Measurement Specialties, Inc., Norristown, PA, http://www.msiusa.com
7. Fraden J (1985) Cardio-respiration transducer. US Patent 4509527, April 9.

8. Del Prete Z, Monteleone L, Steindler R (2001) A novel pressure array sensor based on contact resistance variation: metrological properties. Rev Sci Instrum 72(3):1548–1558
9. Yates B (1991) A keyboard controlled joystick using force sensing resistor. Sensors 39
10. Huff MA, Nikolich AD, Schmidt MA (1991) A threshold pressure switch utilizing plastic deformation of silicon. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of technical papers, pp. 177–180, ©IEEE
11. Jiang JC, White RC, Allen PK (1991) Microcavity vacuum tube pressure sensor for robotic tactile sensing. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of technical papers, pp. 239–240, IEEE
12. Brodie I (1989) Physical considerations in vacuum microelectronics devices. IEEE Trans Electron Devices 36:2641
13. Touch screen controller technology and application trends. Fujitsu Technology Backgrounder. Fujitsu Michroelectronics America, Inc.
14. Draeger C, Fink M (1997) One-channel time reversal of elastic waves in a chaotic 2D-silicon cavity. Phys Rev Lett 79:407–410
15. Benes E, Grösechl M, Burger W, Schmid M (1995) Sensors based on piezoelectric resonators. Sens Actuators A 48:1–2
16. Karrer E, Leach J (1977) A low range quartz pressure transducer. ISA Trans 16:90–98
17. Corbett JP (1991) Quatz steady-state force and pressure sensor. In: Sensors Expo West proceedings, paper 304A-1, 1991. © Helmers Publishing, Inc., Peterborough, NH
18. Kirman RG, Langdon RM (1986) Force sensors. US Patent 4,594,898, June 17.

Chapter 10
Pressure Sensors

To learn something new,
first, you must know something old.

– My physics teacher

10.1 Concepts of Pressure

The concept of pressure was primarily based on the pioneering work of Evangelista Torricelli who for a short time was a student of Galileo [1]. During his experiments with mercury filled dishes, in 1643, he realized that the atmosphere exerts pressure on Earth. Another great experimenter Blaise Pascal, in 1647, conducted an experiment with the help of his brother-in-law, Perier, on the top of the mountain Puy de Dome and at its base. He observed that pressure exerted on the column of mercury depends on elevation. He named a mercury-in-vacuum instrument they used in the experiment the barometer. In 1660, Robert Boyle stated his famous relationship: "*The product of the measures of pressure and volume is constant for a given mass of air at fixed temperature*." In 1738, Daniel Bernoulli developed an impact theory of gas pressure to the point where Boyle's law could be deducted analytically. Bernoulli also anticipated the Charles–Gay-Lussac law by stating that pressure is increased by heating gas at a constant volume. For a detailed description of gas and fluid dynamics, a reader should be referred to one of the many books on the fundamentals of physics. Below, we briefly summarize the basics, which are essential for understanding design and use of pressure sensors.

In general terms, matter can be classified into *solids* and *fluids*. The word fluid describes something that can flow. This includes liquids and gases. The distinction between liquids and gases is not quite definite. By varying pressure, it is possible to change liquid into gas and vice versa. It is impossible to apply pressure to a fluid in any direction except normal to its surface. At any angle, except 90°, fluid will just slide over or flow. Therefore, any force applied to fluid is tangential, and the

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_10, © Springer Science+Business Media, LLC 2010

pressure exerted on boundaries is normal to the surface. For a fluid at rest, pressure can be defined as the force F exerted perpendicularly on a unit area A of a boundary surface [2]:

$$p = \frac{dF}{dA}. \tag{10.1}$$

Pressure is basically a mechanical concept that can be fully described in terms of the primary dimensions of mass, length, and time. It is a familiar fact that pressure is strongly influenced by the position within the boundaries; however, at a given position, it is quite independent of direction. We note the expected variations in pressure with elevation

$$dp = -wdh, \tag{10.2}$$

where w is the specific weight of the medium, and h represents the vertical height.

Pressure is unaffected by the shape of the confining boundaries. Thus, a great variety of pressure sensors can be designed without the concern of shape and dimensions. If pressure is applied to one of the sides of the surface confining fluid or gas, pressure is transferred to the entire surface without diminishing in value.

Kinetic theory of gases states that pressure can be viewed as a measure of the total kinetic energy of the molecules "attacking" the surface

$$p = \frac{2}{3}\frac{KE}{V} = \frac{1}{3}\rho C^2 = NRT, \tag{10.3}$$

where KE is the kinetic energy, V is the volume, C^2 is an average value of the square of the molecular velocities, ρ is the density, N is the number of molecules per unit volume, R is a specific gas constant, and T is the absolute temperature.

Equation (10.3) suggests that pressure and density of compressible fluids (gases) are linearly related. The increase in pressure results in the proportional increase in density. For example, at 0°C and 1 atm. pressure, air has a density of 1.3 kg/m^3, while at the same temperature and 50 atm. pressure its density is 65 kg/m^3, which is 50 times higher. On the contrary, for liquids the density varies very little over ranges of pressure and temperature. For instance, water at 0°C and 1 atm. has a density of 1,000 kg/m^3, while at 0°C and 50 atm., its density is 1,002 kg/m^3, and at 100°C and 1 atm. its density is 958 kg/m^3.

If gas pressure is above or below the pressure of ambient air, we speak about overpressure or partial vacuum. Pressure is called relative when it is measured with respect to ambient pressure. It is called absolute when it is measured with respect to a vacuum at zero pressure. The pressure of a medium may be static when it is referred to fluid at rest, or dynamic when it is referred to kinetic energy of a moving fluid.

10.2 Units of Pressure

The SI unit of pressure is the pascal: 1 Pa = 1 N/m^2. That is, one pascal is equal to 1 N force uniformly distributed over 1 m^2 of surface. Sometimes, in technical systems, the atmosphere is used, which is denoted 1 atm. One atmosphere is the pressure exerted on 1 cm^2 by a column of water having height of 1 m at a temperature of +4°C and normal gravitational acceleration. 1 Pa may be converted into other units by the use of following relationships (see also Table A.4):

$$1\text{Pa} = 1.45 \times 10^{-4}\text{lb/in}^2 = 9.869 \times 10^{-6}\text{atm.} = 7.5 \times 10^{-4}\text{cmHg.}$$

One Pa is quite a low pressure. For a practical estimation, it is useful to remember that 0.1 mmH_2O is roughly equal to 1 Pa. In industry, another unit of pressure is often used. It is defined as pressure exerted by 1 mm column of mercury at 0°C at normal atmospheric pressure and normal gravity. This unit is named after Torricelli and is called the Torr:

$$1 \text{ Torr} = 1 \text{ mmHg}$$

The ideal pressure of the Earth atmosphere is 760 Torr (mmHg) and is called the physical atmosphere

$$1 \text{ atm.} = 760 \text{ Torr} = 101{,}325 \text{ Pa.}$$

For example, in medicine the arterial blood pressure (ABP) traditionally is measured in mmHg and for a healthy person a typical ABP is about 120/70 mmHg, where the first number is the systolic pressure (at heart contraction), while the second number is a diastolic pressure (at heart relaxation). These pressures can be expressed as 0.158/0.092 atm., meaning that in the arteries, blood flows at pressures higher than the ambient atmospheric pressure by these numbers.

The U.S. Customary System of units defines pressure as a pound per square inch (lbs/sq) or psi. Conversion into SI systems is the following:

$$1 \text{ psi} = 6.89 \times 10^3 \text{ Pa} = 0.0703 \text{ atm.}$$

A pressure sensor is a complex sensor. That is, more than one step of the energy conversion is required before pressure can be finally converted into an electrical signal. Thus, the operating principle of many pressure sensors is based on the conversion of a result of the pressure exertion on a sensitive element having a defined surface area. In turn, the element is displaced or deformed. As a result, a pressure measurement may be reduced to a measurement of a displacement or force, which results from a displacement. We recommend that the readers first familiarize themselves with the displacement sensors covered in Chap. 7 and force sensors of Chap. 9.

10.3 Mercury Pressure Sensor

A simple yet efficient sensor is based on communicating vessels principle (Fig. 10.1). Its prime use is for the measurement of gas pressure. A U-shaped wire is immersed into mercury, which shorts its resistance in proportion with the height of mercury in each column. The resistors are connected into a Wheatstone bridge circuit that remains in balance as long as the differential pressure in the tube is zero. Pressure is applied to one of the arms of the tube and disbalances the bridge, which results in the output signal. The higher the pressure in the left tube, the higher the resistance of the corresponding arm and the lower the resistance of the opposite arm. The output voltage is proportional to a difference in resistances ΔR of the wire arms that are not shunted by mercury:

$$V_{out} = V\frac{\Delta R}{R} = V\beta\Delta p. \tag{10.4}$$

The sensor can be directly calibrated in units of Torr. While being simple, this sensor suffers from several drawbacks such as necessity of precision leveling, susceptibility to shocks and vibration, large size, and contamination of gas by mercury vapors.[1]

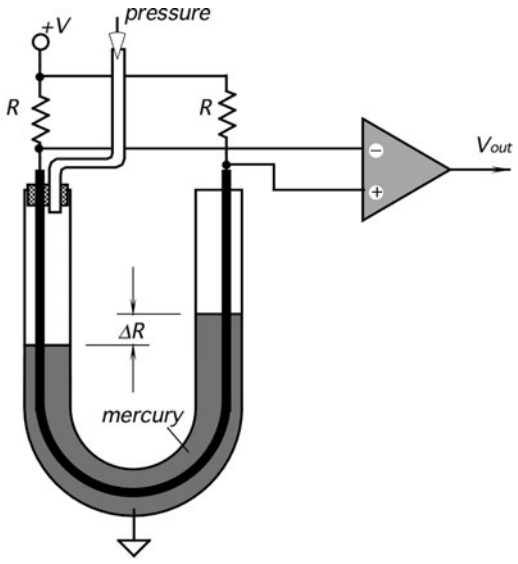

Fig. 10.1 Mercury-filled, U-shaped sensor for measuring gas pressure

[1]Note that this sensor can be used as an inclination sensor when pressures at both sides of the tube are equal.

10.4 Bellows, Membranes, and Thin plates

As it was mentioned earlier, in most cases the pressure sensors contain deformable elements whose deformations or movements are measured and converted by the displacement sensors into electrical signals representative of the pressure values. In pressure sensors, this deformable or sensing element is a mechanical device that undergoes structural changes under strain resulting from pressure. Historically, such devices were bourdon tubes (C-shaped, twisted, and helical), corrugated [3] and catenary diaphragms, capsules, bellows, barrel tubes, and other components whose shape was changing under pressure.

A bellows (Fig. 10.2a) is intended for the conversion of pressure into a linear displacement, which can be measured by an appropriate sensor. Thus, bellows performs a first step in the complex conversion of pressure into an electrical signal. The bellows is characterized by a relatively large surface area and, therefore, by a large displacement at low pressures [23]. The stiffness of seamless metallic bellows is proportional to the Young's modulus of the material and inversely proportional to the outside diameter and to the number of convolutions of the bellows. Stiffness also increases with roughly the third power of the wall thickness.

A popular example of pressure conversion into a linear deflection is a diaphragm in an aneroid barometer (Fig. 10.2b) . A deflecting device always forms at least one wall of a pressure chamber and is coupled to a strain sensor (for instance, a strain gauge like the one shown in Fig. 9.2), which converts deflection into electrical signals by means of piezoresistivity. Nowadays, a great majority of pressure sensors are fabricated with silicon membranes by using a micromachining technology.

A membrane is a thin diaphragm under radial tension S, which is measured in N/m (Fig. 10.3b). The stiffness to bending forces can be neglected as the thickness of the membrane is much smaller when compared with its radius (at least 200 times smaller). When pressure is applied to one side of a membrane, it shapes spherically,

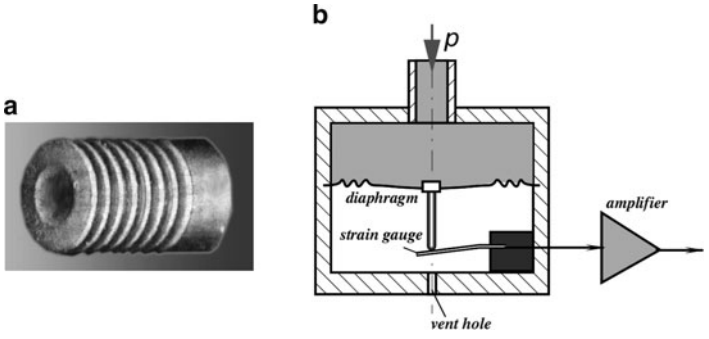

Fig. 10.2 Steel bellows for a pressure transducer (**a**) and metal corrugated diaphragm for the conversion of pressure into linear deflection (**b**)

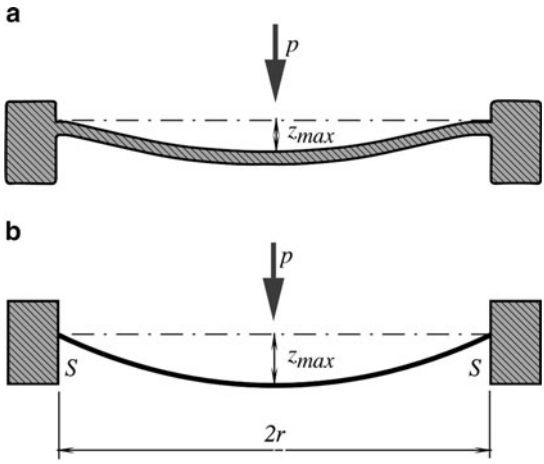

Fig. 10.3 Thin plate (**a**) and membrane (**b**) under pressure p

like a soap bubble. At low-pressure p differences across the membrane, the center deflection z_{max} and the stress σ_{max} are quasilinear functions of pressure:[2]

$$z_{max} = \frac{r^2 p}{4S}, \tag{10.5}$$

$$\sigma_{max} \approx \frac{S}{g}, \tag{10.6}$$

where r is the membrane radius and g is the thickness. Stress is generally uniform over the membrane area.

For the membrane, the lowest natural frequency can be calculated from [4]

$$f_0 = \frac{1.2}{\pi r}\sqrt{\frac{S}{\rho g}}, \tag{10.7}$$

where ρ is the membrane material density.

If the thickness of the membrane is not negligibly small (r/g ratio is 100 or less), the membrane is no longer a "membrane," and it is called a thin plate (Fig. 10.3a). If the plate is compressed between some kinds of clamping rings, it exhibits a noticeable hysteresis due to friction between the thin plate and the clamping rings. A much better arrangement is a one-piece structure where the plate and the supporting components are fabricated of a single bulk of material.

For a plate, the maximum deflection is also linearly related to pressure

[2]Stress is measured in N/m^2.

$$z_{max} = \frac{3(1-v^2)r^4 p}{16Eg^3}, \tag{10.8}$$

where E is Young's modulus (N/m^2) and v is Poisson's ratio. The maximum stress at the circumference is also a linear function of pressure:

$$\sigma_{max} \approx \frac{3r^2 p}{4g^2}. \tag{10.9}$$

The above equations suggest that a pressure sensor can be designed by exploiting the membrane and thin plate deflections. The next question is: What physical effect to use for the conversion of deflection into an electrical signal? There are several options that we discuss in the following paragraphs.

10.5 Piezoresistive Sensors

To make a pressure sensor, two essential components are required. They are the plate (membrane) having known area A and a detector that responds to applied force F (10.1). Both these components can be fabricated of silicon. A silicon-diaphragm pressure sensor consists of a thin silicon diaphragm as an elastic material [5] and piezoresistive gauge resistors made by diffusive impurities into the diaphragm. Thanks to single crystal silicon superior elastic characteristics, virtually no creep and no hysteresis occur, even under strong static pressure. The gauge factor of silicon is many times stronger than that of thin metal conductors [6]. It is customary to fabricate strain gauge resistors connected as the Wheatstone bridge. The full-scale output of such a circuit is on the order of several hundred millivolts; thus, a signal conditioner is required for bringing the output to an acceptable format. Further, silicon resistors exhibit quite strong temperature sensitivity; therefore, either the piezoresistors should be temperature compensated or a signal conditioning circuit should include temperature compensation.

When stress is applied to a semiconductor resistor having initial resistance R, piezoresistive effect results in change in the resistance ΔR [7]:

$$\frac{\Delta R}{R} = \pi_l \sigma_l + \pi_t \sigma_t, \tag{10.10}$$

where π_l and π_t are the piezoresistive coefficients in a longitudinal and transverse directions, respectively. Stresses in longitudinal and transverse directions are designated σ_l and σ_t. The π-coefficients depend on the orientation of resistors on the silicon crystal. Thus, for p-type diffused resistor arranged in the <110> direction or an n-type silicon square diaphragm with (100) surface orientation as shown in Fig. 10.4, the coefficients are approximately denoted as [7]

$$\pi_l = -\pi_t = \frac{1}{2}\pi_{44}. \tag{10.11}$$

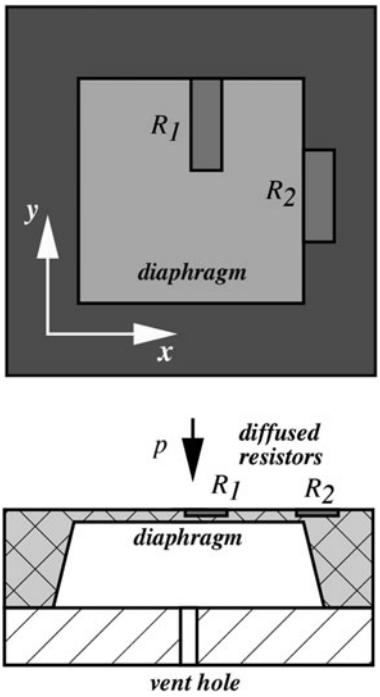

Fig. 10.4 Position of piezoresistors on a silicon diaphragm

A change in resistivity is proportional to applied stress and, subsequently, to applied pressure. The resistors positioned on the diaphragm in such a manner as to have the longitudinal and transverse coefficients of the opposite polarities, therefore, resistors change in the opposite directions:

$$\frac{\Delta R_1}{R_1} = -\frac{\Delta R_2}{R_2} = \frac{1}{2}\pi_{44}\left(\sigma_{1y} - \sigma_{1x}\right). \tag{10.12}$$

When connecting R_1 and R_2 in a half-bridge circuit and exciting the bridge with E, the output voltage is

$$V_{out} = \frac{1}{4}E\pi_{44}\left(\sigma_{1y} - \sigma_{1x}\right). \tag{10.13}$$

As a result, pressure sensitivity a_p and temperature sensitivity of the circuit b_T can be found by taking partial derivatives:

$$a_p = \frac{1}{E}\frac{\partial V_{out}}{\partial p} = \frac{\pi_{44}}{4}\frac{\partial\left(\sigma_{1y} - \sigma_{1x}\right)}{\partial p}, \tag{10.14}$$

$$b_T = \frac{1}{a_p}\frac{\partial a_p}{\partial T} = \frac{1}{\pi_{44}}\frac{\partial \pi_{44}}{\partial T}. \tag{10.15}$$

Since $\partial\pi_{44}/\partial T$ has a negative value, the temperature coefficient of sensitivity is negative, that is, sensitivity decreases at higher temperatures.

There are several methods of fabrication, which can be used for the silicon pressure sensor processing. In one method [8], the starting material is *n*-type silicon substrate with (100) surface orientation. Piezoresistors with 3×10^{18} cm^{-3} surface-impurity concentration are fabricated using a boron ion implantation. One of them (R_1) is parallel, and the other is perpendicular to the <110> diaphragm orientation. Other peripheral components, like resistors and pn-junctions used for temperature compensation, are also fabricated during the same implantation process as that for the piezoresistors. They are positioned in a thick-rim area surrounding the diaphragm. Thus, they are insensitive to pressure applied to the diaphragm.

Another approach of stress sensing was used in Motorola MPX pressure sensor chip shown in Fig. 10.5. The piezoresistive element, which constitutes a strain gauge, is ion implanted on a thin silicon diaphragm. Excitation current is passed longitudinally through the resistor's taps 1 and 3, and the pressure that stresses the diaphragm is applied at a right angle to the current flow. The stress establishes a transverse electric field in the resistor that is sensed as voltage at taps 2 and 4. The single-element transverse voltage strain gauge can be viewed as the mechanical analog of a Hall effect device (see Sect. 3.8). Using a single element eliminates the need to closely match the four stress and temperature sensitive resistors that form a Wheatstone bridge design. At the same time, it greatly simplifies the additional circuitry necessary to accomplish calibration and temperature compensation. Nevertheless, the single element strain gauge electrically is analogous to the bridge circuit. Its balance (offset) does not depend on matched resistors, as it would be in a conventional bridge, but instead on how well the transverse voltage taps are aligned.

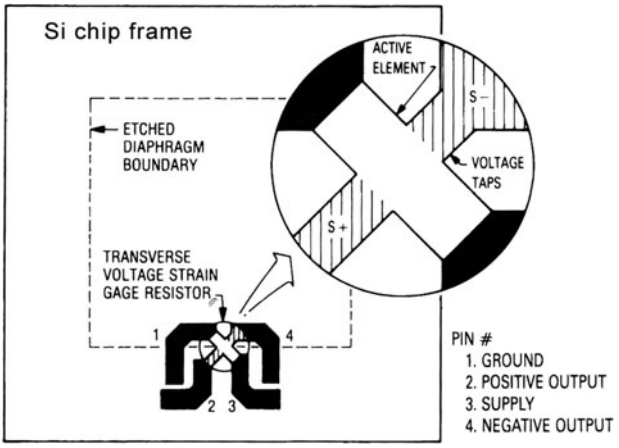

Fig. 10.5 Basic uncompensated piezoresistive element of Motorola MPX pressure sensor (Copyright of Motorola, Inc. Used with permission)

A thin diaphragm with 1 mm^2 area size may be formed by using one of the commonly used silicon etching solutions, for instance, hydrazine–water ($N_2H_4 \cdot H_2O$) anisotropic etchant. A SiO_2 or Si_3N_4 layer serves as an etch mask and the protective layer on the bottom side of the wafer. The etching rate is about 1.7 μm/min at 90°C in reflux solution. The final diaphragm thickness is achieved at about 30 μm.

Another method of diaphragm fabrication is based on the so-called silicon fusion bonding (SFB) where single crystal silicon wafers can be reliably bonded with near-perfect interfaces without the use of intermediate layers [9]. This technique allows the making of very small sensors that find use in catheter-tip transducers for medical in vivo measurements. Total chip area may be as much as eight times smaller than that of the conventional silicon-diaphragm pressure sensor. The sensor consists of two parts – the bottom and the top wafers (Fig. 10.6a). The bottom constraint wafer (substrate) is first anisotropically etched with a square hole that has desirable dimensions of the diaphragm. The bottom wafer has thickness about 0.5 mm, and the diaphragm has side dimensions of 250 μm, so the anisotropic etch forms a pyramidal cavity with a depth of about 175 μm. The next step is SFB bonding to a top wafer consisting of a *p*-type substrate with an *n*-type epi layer. The thickness of the epi layer corresponds to the desired final thickness of the diaphragm. Then, the bulk of the top wafer is removed by a controlled-etch process, leaving a bonded-on single crystal layer of silicon that forms the sensor's diaphragm. Next, resistors are ion implanted and contact vias are etched. In the final step, the constrain wafer is ground and polished back to the desired thickness of the device, about 140 μm. Despite the act that the dimensions of the SFB chip are about half of those of the conventional chip, their pressure sensitivities are identical. A comparison of conventional and SFB technology is shown in Fig. 10.6b. For the same diaphragm dimensions and the same overall thickness of the chip, the SFB device is about 50% smaller.

A diaphragm (membrane) of a piezoresistive sensor in many sensors usually is very thin, on the order of 1 μm; thus its mechanical properties are a limiting factor

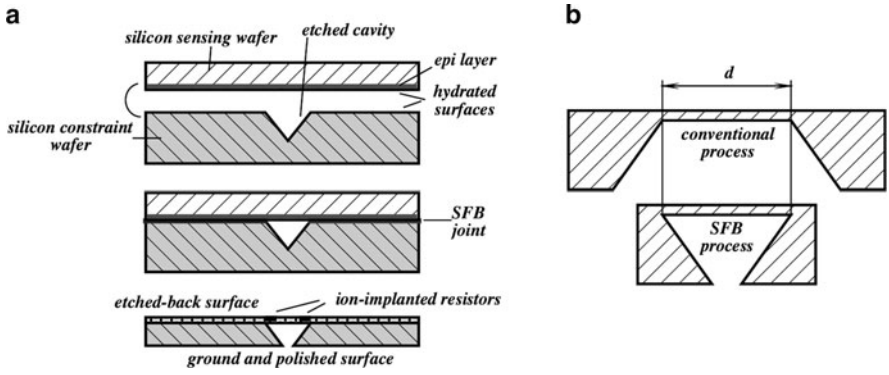

Fig. 10.6 Silicon membrane fabrication production steps of silicon fusion bonding method (**a**); comparison of an SFB chip size with a conventionally fabricated diaphragm (**b**)

for the maximum applied pressures. In applications where pressures are very high, the silicon diaphragm is just too weak to be directly subjected to such pressures. Thus, the force applied to the silicon diaphragm should be scaled down by the use of an intermediate pressure plate. For example, in automotive industry for measuring pressure in combustion engines where temperature reaches 2,000°C and pressures may exceed 200 bar, a special sensor housing with a scaling pressure plate may be employed. Such a housing should scale down pressure and protect the chip from a harsh environment. Figure 10.7 illustrates a housing where pressure sensitive chip with a micromachined Si diaphragm is positioned above the steel plate. High pressures flex the steel plate with a relatively small displacement in the center section that carries the boss. The boss is mechanically coupled to the Si diaphragm and flexes it upwards causing disbalance of the piezoresistive bridge.

Pressure sensors are usually available in three basic configurations that permit measurement of absolute, differential, and gauge pressures. Absolute pressure, such as barometric pressure, is measured with respect to a reference vacuum chamber. The chamber may be either external, or it can be built directly into the sensor (Fig. 10.8a). A differential pressure, such as the pressure drop in a pressure-differential flowmeter, is measured by applying pressure to the opposite sides of the diaphragm simultaneously. Gauge pressure is measured with respect to some kind of reference pressure. An example is ABP measurement, which is done with

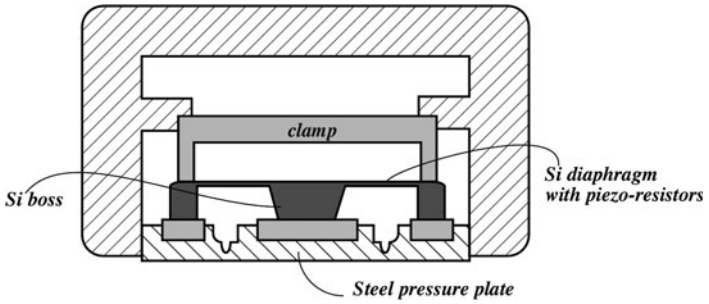

Fig. 10.7 Piezoresistive chip inside the steel enclosure for measuring high pressures

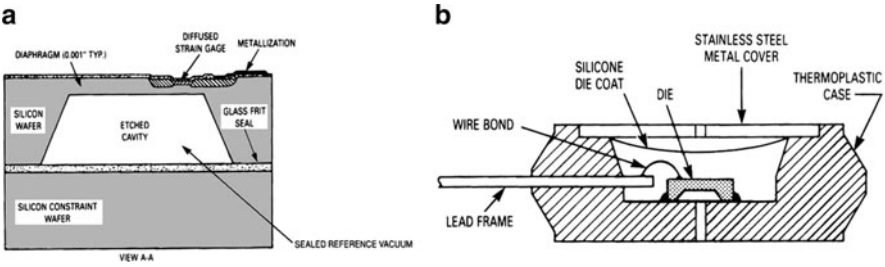

Fig. 10.8 Absolute (**a**) and differential (**b**) pressure sensor packagings (Copyright of Motorola, Inc. Used with permission)

respect to atmospheric pressure. Thus, gauge pressure is a special case of a differential pressure. Diaphragm and strain gauge designs are the same for all three configurations, while the packaging makes them different. For example, to make a differential or gauge sensor, a silicon die is positioned inside the chamber that has two openings at both sides of the die (Fig. 10.8b). To protect them from harsh environment, the interior of the housing is filled with a silicone gel, which isolates the die surface and wire bonds, while allowing the pressure signal to be coupled to the silicon diaphragm. A differential sensor may be incorporated into various porting holders (Fig. 10.9). Certain applications, such as a hot water hammer, corrosive fluids, and load cells, require physical isolation and hydraulic coupling to the chip-carrier package. It can be done with additional diaphragms, plates, and bellows as exemplified in Fig. 10.9. In either case, silicon oil, such as Dow Corning DS200, can be used to fill the air cavity so that system frequency response is not degraded.

All silicon-based sensors are characterized by temperature dependence. Temperature coefficient of sensitivity b_T as defined by (10.15) is usually negative, and for the accurate pressure sensing, it must be compensated for. Without the compensation, sensor's output voltage may look like the one shown in Fig. 10.10a for three different temperatures.

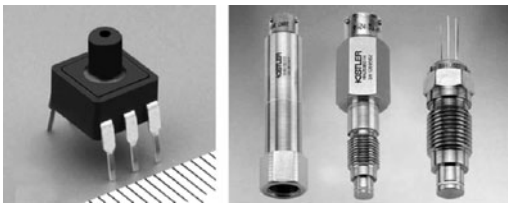

Fig. 10.9 Examples of pressure sensor packagings

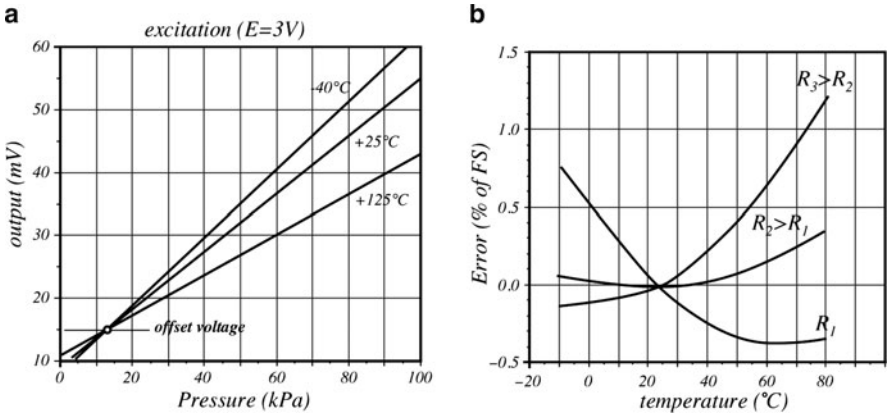

Fig. 10.10 Temperature characteristics of a piezoresistive pressure sensor. Transfer function at three different temperatures (**a**) and full-scale errors for three values of compensating resistors (**b**)

In many applications, a simple yet efficient temperature compensation can be accomplished by adding to the sensor either a series or a parallel temperature-stable resistor. By selecting an appropriate value of the resistor, the sensor's output can be tailored to the desirable operating range (Fig. 10.10b). Whenever a better temperature correction over a broad range is required, more complex compensation circuits with temperature detectors can be employed. A viable alternative is a software compensation where the temperature of the pressure transducer is measured by an imbedded temperature sensor. Data from both the pressure and temperature sensors are relayed to the processing circuit where numerical compensation is digitally performed. But the best solution is still designing a temperature-compensated Si bridge inside the sensor.

10.6 Capacitive Sensors

A silicon diaphragm can be used with another pressure-to-electric output conversion process: a capacitive sensor. In a capacitive pressure sensor, the diaphragm displacement modulates capacitance with respect to the reference plate (backplate). This conversion is especially effective for the low-pressure sensors. An entire sensor can be fabricated from a solid piece of silicon, thus maximizing its operational stability. The diaphragm can be designed to produce up to 25% capacitance change over the full range, which makes these sensors candidates for direct digitization (see Sect. 5.6). While a piezoresistive diaphragm should be designed to maximize stress at its edges, the capacitive diaphragm utilizes a displacement of its central portion. These diaphragms can be protected against overpressure by including mechanical stops close to either side of the diaphragm (for a differential pressure sensor). Unfortunately, in the piezoresistive diaphragms, the same protection is not quite effective because of small operational displacements. As a result, the piezoresistive sensors typically have burst pressures of about ten times the full-scale rating, while capacitive sensors with overpressure stops can handle a thousand times the rated full-scale pressure. This is especially important for the low-pressure applications, where relatively high-pressure pulses can occasionally occur.

While designing a capacitive pressure sensor, for good linearity, it is important to maintain flatness of the diaphragm. Traditionally, these sensors are linear only over the displacements that are much less then their thickness. One way to improve the linear range is to make a diaphragm with groves and corrugations by applying a micromachining technology. Planar diaphragms are generally considered more sensitive than the corrugated diaphragms with the same size and thickness. However, in the presence of the in-plane tensile stresses, the corrugations serve to release some of the stresses, thus resulting in better sensitivity and linearity (Fig. 10.11a). An effective way of improving a linearity and sensitivity is to combine a bellows with a flat plate (Fig. 10.11b).

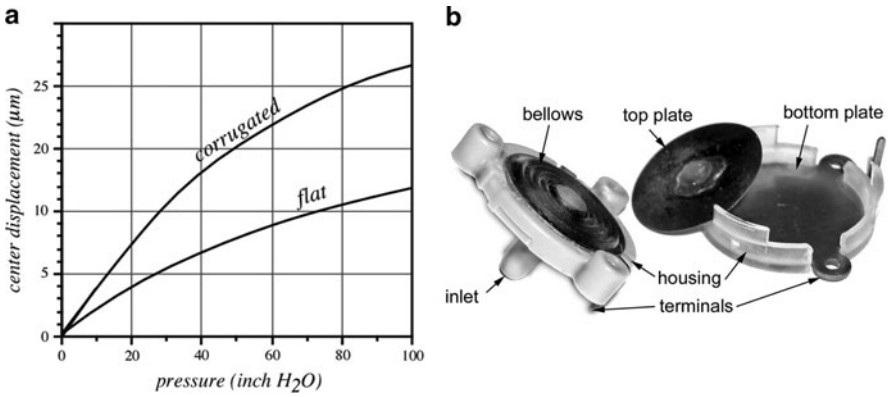

Fig. 10.11 Central deflection of flat and corrugated diaphragms of the same sizes under the in-plate tensile stresses (**a**); disassembled capacitive sensor with a bellows (**b**)

10.7 VRP Sensors

When measuring small pressures, deflection of a thin plate or a diaphragm can be very small. In fact, it can be so small that the use of strain gauges attached to or embedded into the diaphragm becomes impractical due to the low-output signal. One possible answer to the problem may be a capacitive sensor where a diaphragm deflection is measured by its relative position to a reference base rather than by the internal strain in the material. Such sensors were described earlier in this chapter. Another solution that is especially useful for very low pressures is a magnetic sensor. A variable-reluctance pressure (VRP) sensor uses a magnetically conductive diaphragm to modulate the magnetic resistance of a differential transformer . The operation of the sensor is very close to that of the magnetic proximity detectors described in Sect. 7.3. Figure 10.12a illustrates a basic idea behind the magnetic flux modulation. The assembly of an E-shaped core and a coil produces a magnetic flux whose field lines travel through the core, the air gap, and the diaphragm. The permeability of the E-core magnetic material is at least 1,000 times higher than that of the air gap [10], and, subsequently, its magnetic resistance is lower than the resistance of air. Since the magnetic resistance of the air gap is much higher than the resistance of the core, it is the gap that determines the inductance of the core–coil assembly. When the diaphragm deflects, the air gap increases or decreases depending on the direction of a deflection, thus causing the modulation of the inductance.

To fabricate a pressure sensor, a magnetically permeable diaphragm is sandwiched between two halves of the shell (Fig. 10.13). Each half incorporates a E-core/coil assembly. The coil s are encapsulated in a hard compound to maintain maximum stability under even very high pressure. Thin pressure cavities are formed at both sides of the diaphragm. The thickness of the diaphragm defines a full-scale operating range; however, under most of circumstances, total deflection

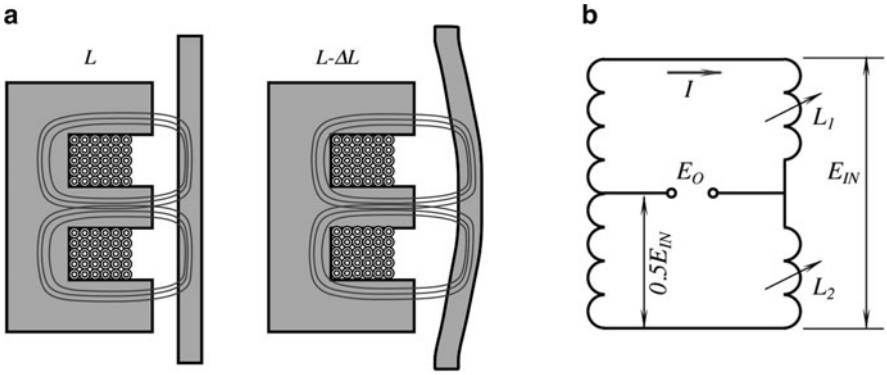

Fig. 10.12 Variable-reluctance pressure sensor. Basic principle of operation (**a**) and equivalent circuit (**b**)

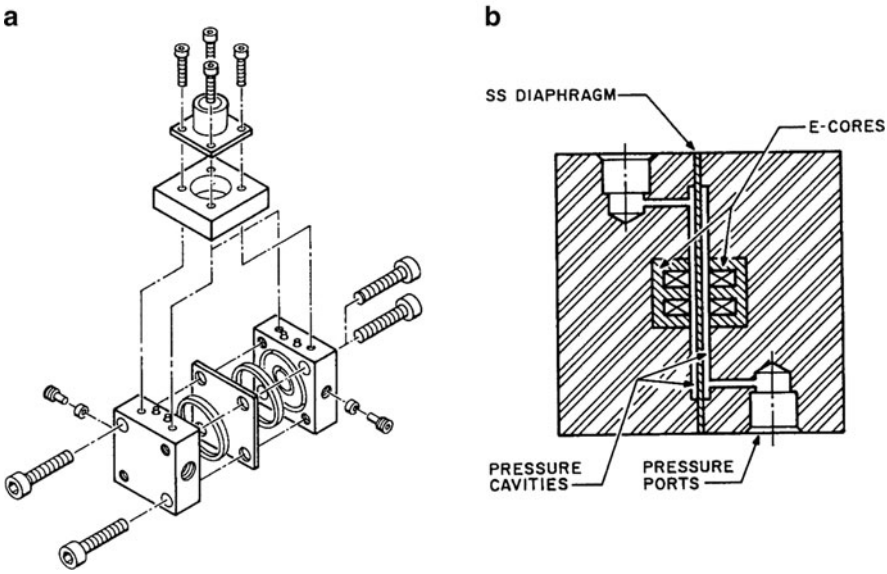

Fig. 10.13 Construction of a VRP sensor for low-pressure measurements. Assembly of the sensor (**a**) and double E-core at both sides of the cavity (**b**)

does not exceed 25–30 μm, which makes this device very sensitive to low pressures. Further, due to thin pressure cavities, the membrane is physically prevented from excessive deflection under the overpressure conditions. This makes VRP sensors inherently safe devices. When excited by an ac current, a magnetic flux is produced in each core and the air gaps by the diaphragm. Thus, the sensors contain two inductances and can, therefore, be thought of as half of a variable-reluctance bridge where each inductance forms one arm of the bridge (Fig. 10.12b). As a differential

pressure across the diaphragm is applied, the diaphragm deflects, one side decreasing and the other increasing, and the air gap reluctances in the electromagnetic circuit change proportionally to the differential pressure applied. A full-scale pressure on the diaphragm, while being very small, will produce a large output signal that is easily differentiated from noise.

VRP sensor's output is proportional to the reluctance in each arm of the inductive Wheatstone bridge that uses the equivalent inductive reactances $x_{1,2}$ as the active elements. The inductance of a coil is determined by the number of turns and the geometry of the coil. When a magnetically permeable material is introduced into the field flux, it forms a low-resistance path attracting magnetic field . This alters the coil's self-inductance. The inductance of the circuit, and subsequently its reactance, is inversely proportional to the magnetic reluctance, that is, $x_{1,2} = k/d$, where k is a constant and d is the gap size. When the bridge is excited by a carrier current, the output signal across the bridge becomes amplitude-modulated by the applied pressure. The amplitude is proportional to the bridge imbalance, and the phase of the output signal changes with the direction of the imbalance. The ac signal can be demodulated to produce a dc response.

10.8 Optoelectronic Pressure Sensors

When measuring low-level pressures or, contrary, when thick membranes are required to enable a broad dynamic range, a diaphragm displacement may be too small to assure a sufficient resolution and accuracy. Besides, most of piezoresistive sensors, and some capacitive, are quite temperature-sensitive, which requires an additional thermal compensation. An optical readout has several advantages over other technologies, namely, a simple encapsulation, small temperature effects, and high resolution and accuracy. Especially promising are the optoelectronic sensors operating with the light interference phenomena [11]. Such sensors use a Fabry–Perot (FP) principle of measuring small displacements as covered in more detail in Sect. 7.4.4. A simplified circuit of one such a sensor is shown in Fig. 10.14.

The sensor consists of the following essential components: A passive optical pressure chip with a membrane etched in silicon, a light-emitting diode (LED), and a detector chip [12]. A pressure chip is similar to a capacitive pressure sensor as described earlier in this chapter, except that a capacitor is replaced by an optical cavity forming a Fabry–Perot (FP) interferometer [13] measuring the deflection of the diaphragm. A back-etched, single-crystal diaphragm on a silicon chip is covered with a thin metallic layer, and a glass plate with a metallic layer on its backside. The glass is separated from the silicon chip by two spacers at a distance w. Two metallic layers form a variable-gap FP interferometer with a pressure-sensitive movable mirror (on the membrane) and a plain-parallel stationary fixed half-transparent mirror (on the glass). A detector chip contains three *pn*-junction photodiodes. Two of them are covered with integrated optical FP filters of slightly different thicknesses. The filters are formed as first surface silicon mirrors coated with

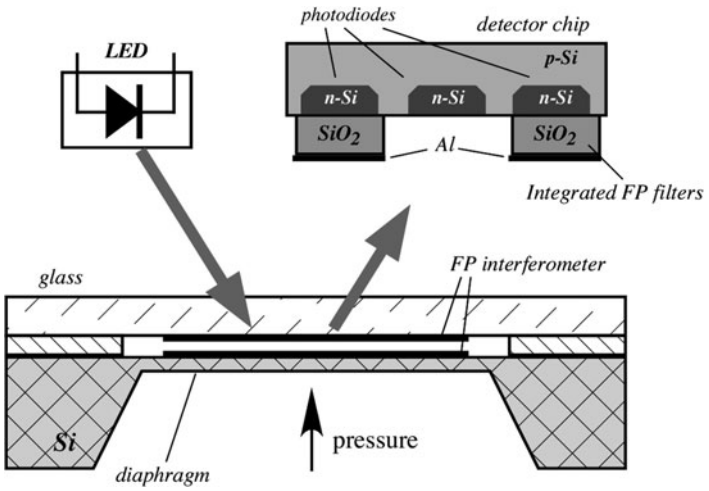

Fig. 10.14 Schematic of an optoelectronic pressure sensor operating on the interference phenomenon (adapted from [12])

a layer of SiO_2 and thin metal (Al) mirrors on their surfaces. An operating principle of the sensor is based on the measurement of a wavelength modulation of the reflected and transmitted light depending on the width of the FP cavity. The reflection and transmission from the cavity is almost a periodic function in the inverse wavelength, $1/\lambda$, of the light with a period equal to $1/2w$. Since w is a linear function of the applied pressure, the reflected light is wavelength modulated.

The detector chip works as a demodulator and generates electrical signals representing the applied pressure. It performs an optical comparison of the sensing cavity of the pressure sensor with a virtual cavity formed by the height difference between two FP filters. If both cavities are the same, the detector generates the maximum photocurrent, and, when the pressure changes, the photocurrent is cosine modulated with a period defined by a half the mean wavelength of the light source. The photodiode without the FP filter serves as a reference diode, which monitors the total light intensity arriving at the detector. Its output signal is used for the ratiometric processing of the information. Since the output of the sensor is inherently nonlinear, a linearization by a microprocessor is generally required. Similar optical pressure sensors can be designed with fiberoptics, which makes them especially useful for remote sensing where radio frequency interferences present serious problem [14].

10.9 Indirect Pressure Sensor

For measuring very small pressure variations, on the order of few pascals, the diaphragm-based pressure sensors are not really efficient because it is very difficult to make a very thin membrane (diaphragm) that would be required to respond to

small pressures.[3] Even if such a membrane is constructed, it will be fragile and could be easily damaged by an accidental overpressure. Thus, other methods of measuring small pressures were devised.

Let us assume that we have a large enclosed tank filled with air under low-relative static pressure p_1. If we attach a small bleed tube to the interior of the tank and open the other end of the tube to atmosphere having the atmospheric pressure p, air will be flowing out of the tank to the atmosphere through the bleed tube. Naturally, the air will flow backwards – from atmosphere into the tank – if p_1 is lower than atmosphere. Since the tank is large and the tube is small, the air flow rate inside the bleed tube will be considered constant v_2. The pressure differential (tank minus atmosphere) can be derived from the Bernoulli equation (see Sect. 11.2) as

$$\Delta p = p_1 - p_2 = b\rho v_2^2, \tag{10.16}$$

where ρ is the air density, v is the mass flow rate, and b is the scaling coefficient that among other factors depends on size of the bleed tube. Note that the air density is proportional to the average pressure that makes the pressure differential nearly independent on the absolute level of pressure. Equation (10.16) shows that the pressure differential can be expressed in terms of flow rate. Thus, we can monitor gas pressure in the tank without a conventional pressure sensor, but indirectly by measuring the velocity (flow rate) of the outflowing or inflowing gas. It is the basis of a differential pressure sensor that employs a flowmeter [15]. This sensor allows measuring the low-pressure differentials on the order of 0.1 Torr and even smaller. Because of the need for a bleed tube, this method is primarily useful for monitoring pressure gradients in dynamic gas systems, such as HVAC where gas is moved by a blower. In these cases, depending on the orientation of the bleed tube opening, the monitored pressure will be either the static pressure or the stagnation (total) pressure that includes the dynamic pressure of the flowing gas, just like in the Pitot tube.[4]

Figure 10.15 shows a practical implementation of the differential pressure sensor with a bleed tube and a flowmeter. An air duct of a HVAC system conducts flow of air having a static pressure p_1. The bleed tube is inserted into the air duct perpendicular to the vector of air flow; thus, the open end is exposed only to the static pressure p_1. The other side of the tube is open to the atmosphere having pressure p_1. According to (10.16), the pressure differential across the tube will cause air flow through the tube. The bleed tube has a built-in mass flowmeter probe that measures the flow rate v. The pressure gradient can be calculated from (10.16). Note that value b needs a calibration. The simplest and most efficient flowmeter for use in this device is a gas thermoanemometer that is described in Sect. 11.3.

[3] Difficult but not impossible. Thin diaphragms were developed for the vacuum sensors [21], albeit they are expensive and very delicate.

[4] Refer to description of Pitot tubes elsewhere, for example in [16].

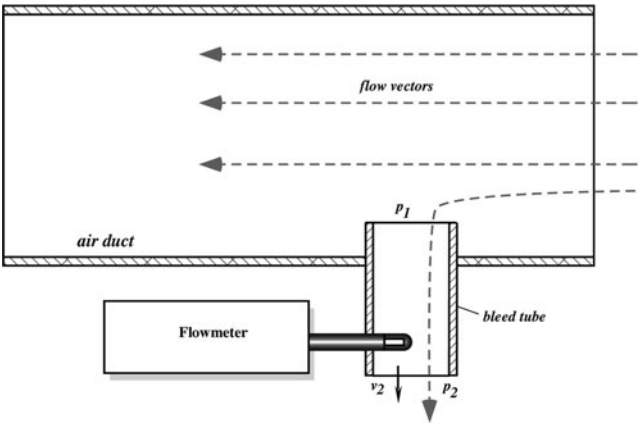

Fig. 10.15 Flowmeter as a differential pressure sensor (adapted from [15])

10.10 Vacuum Sensors

Measurement of extremely low pressures is important for processing of the microelectronic wafers, optical components, chemistry, and other industrial applications [22]. It also vital for the scientific studies, for instance, in the space exploration. In general, vacuum means pressure below atmospheric, but usually the term is used with respect to a near absence of gas pressure. True vacuum is never attained. Even the intrastellar space is not entirely free of matter.

Vacuum can be measured as negative pressure comparing to the atmospheric pressure by conventional pressure sensors, yet this is not quite efficient. Conventional pressure sensors do not resolve extremely low concentrations of gas due to poor signal-to-noise ratio. While a pressure sensor in most cases employs some kind of membrane and a displacement (deflection) transducer, special vacuum sensors operate on different principles. They rely on certain physical properties of gaseous molecules that are related to the number of such molecules per volume of space. These properties may be a thermal conductivity, viscosity, ionization, and others. Here, we briefly describe some popular sensor designs.

10.10.1 Pirani Gauge

Pirani vacuum gauge is a sensor that measures pressure through thermal conductivity of gas. It is one of the oldest vacuum sensors. The simplest version of the gauge contains a heated plate. The measurement is done by detecting the amount of heat loss from the plate that depends on the gas pressure. Operation of the Pirani gauge is based on pioneering works by Von Smoluchowski [17]. He established that when an object is heated, thermal conductivity to the surrounding objects is governed by

$$G = G_0 + G_g = G_s + G_r + ak\frac{PP_T}{P + P_T}, \tag{10.17}$$

where G_s is thermal conductivity via the solid supporting elements, G_e is the radiative heat transfer, a is the area of a heated plate, k is a coefficient related to gas properties, and P_T is a transitional pressure that is the maximum pressure that can be measured. Figure 10.16a illustrates different factors that contribute to a thermal loss from a heated plate. If the solid conductive and radiative loss is accounted for, the gas conductivity G_g goes linearly down to absolute vacuum. The trick is to minimize the interfering factors that contribute to G_0. This can be achieved by the use of both the heated plate that is suspended with a minimal thermal contact with the sensor housing and the differential technique that to a large degree cancels the influence of G_0.

There are several designs of the Pirani gauge that are used in vacuum technologies. Some employ two plates with different temperatures and the amount of power spent for heating is the measure of gas pressure. The others use a single plate that measures thermal conductivity of gas by heat loss to the surrounding walls. Temperature measurement is usually done with either a thermocouple or a platinum RTD. Figure 10.17 illustrates one version of the gauge that employs a thermal balance (differential) technique. The sensing chamber is divided into two identical sections where one is filled with gas at a reference pressure, say 1 atm. = 760 Torr and the other is connected to the vacuum that is to be measured. Each chamber contains a heated plate that is supported by the tiny links to minimize a conductive heat transfer through solids. Both chambers are preferably of the same shape, size, and construction so that the conductive and radiative heat loss would be nearly identical. The better the symmetry, the better the cancellation of the spurious

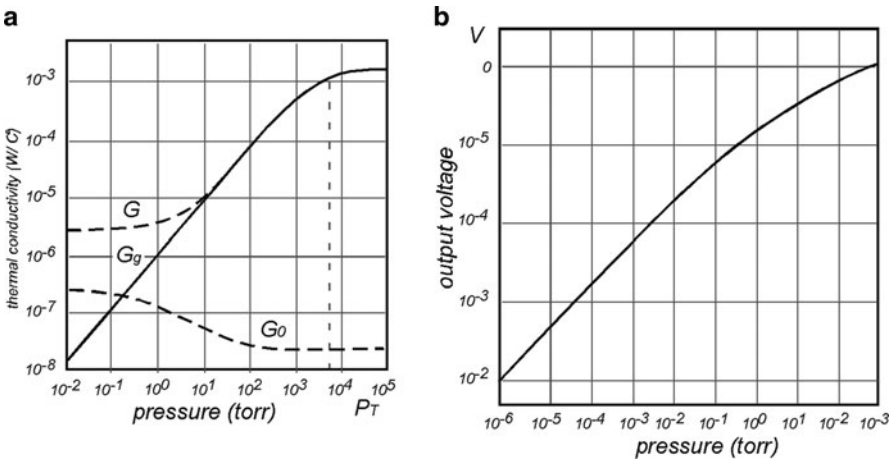

Fig. 10.16 Thermal conductivities from a heated plate (**a**). Transfer function of a Pirani vacuum gauge (**b**)

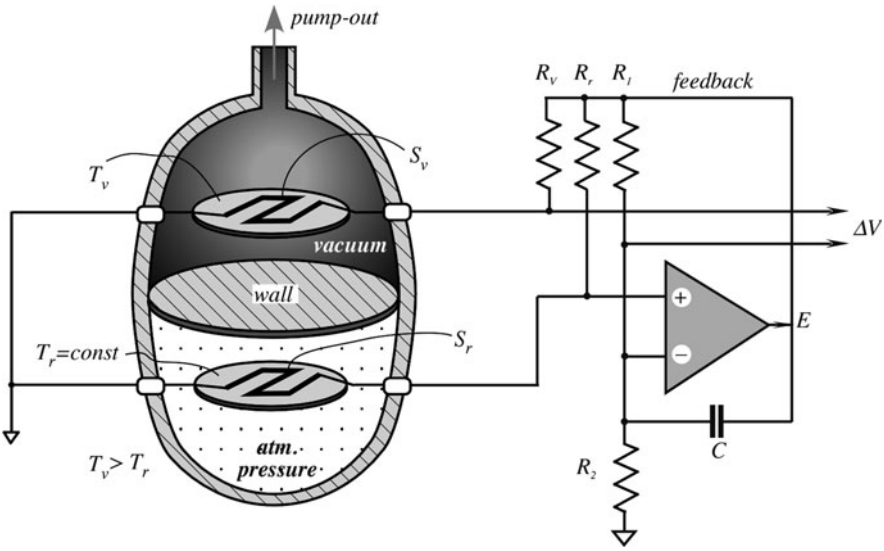

Fig. 10.17 Pirani vacuum gauge with NTC thermistors operating in self-heating mode

thermal conductivity G_0. The heaters on the plates are warmed up by electric current. In this particular design, each heater is a thermistor with a negative temperature coefficient NTC (see Chap. 16). Resistances of the thermistor are equal and relatively low to allow for a Joule self-heating (see Fig. 16.14). The reference thermistor S_r is connected into a self-balancing bridge that also includes fixed resistors R_r, R_1, R_2, and an operational amplifier. The bridge automatically sets temperature of S_r on a constant level T_r that is defined by the bridge resistors and is independent of the ambient temperature. Note that the bridge is balanced by both the negative and positive feedbacks to both bridge arms. Capacitor C keeps the circuit from oscillating. The same voltage E that feeds the reference plate is applied to the thermistor S_V on the sensing plate via $R_v = R_r$. The output voltage ΔV is taken differentially from the sensing thermistor and the bridge. The shape of the transfer function is shown in Fig. 10.16a. A vacuum sensor often operates with gases that may contaminate the sensing plates so the appropriate filters must be employed.

10.10.2 Ionization Gauges

This sensor resembles a vacuum tube that was used as an amplifier in the old-fashioned radio equipment. The ion current between the plate and the filament (Fig. 10.18a) is a nearly linear function of molecular density (pressure) [18]. The vacuum gauge tube has a reversed connection of voltages: The positive high voltage is applied to a grid, while negative lower voltage is connected to the plate. The output

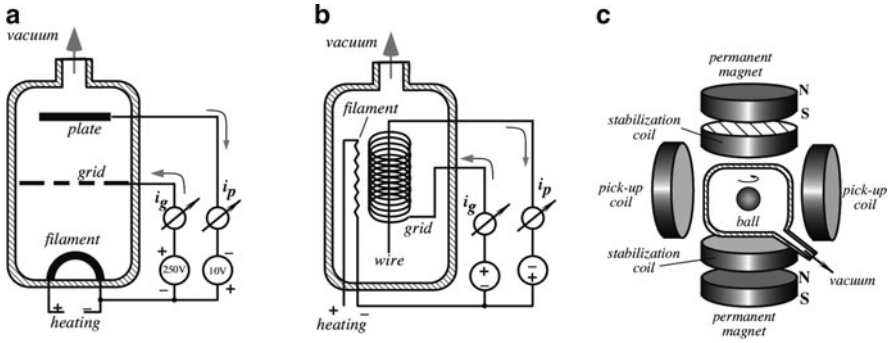

Fig. 10.18 Ionization vacuum gauge (**a**), Bayard-Alpert gauge (**b**), and gas drag gauge (**c**)

is the ion current i_p collected by the plate that is proportional to pressure and the electron current i_g of the grid. Presently, a further improvement of this gauge is the so-called Bayard–Alpert vacuum sensor [19]. It is more sensitive and stable at much lower pressures. Its operating principle is the same as a vacuum tube gauge, except that the geometry is different – the plate is substituted by the wire surrounded by a grid, while the cathode filament is outside (Fig. 10.16b).

10.10.3 Gas Drag Gauge

The gas molecules interact with a moving body. This is the basic idea behind the spinning rotor gauge [20]. In the current implementation of the sensor, a small steel ball having diameter of 4.5 mm is magnetically levitated (Fig. 10.18c) inside a vacuum chamber and spinning with a rate of 400 Hz. The ball magnetic moment induces a signal in a pick-up coil. The gas molecules exert drag on the ball and slow its rate of rotation.

$$\mathrm{P} = \frac{\pi \rho a \bar{c}}{10\sigma_{\mathrm{eff}}} \left(\frac{-\omega' - RD - 2\alpha T}{\omega} \right), \tag{10.18}$$

where ρ and a are the density and radius of the ball, respectively, ω'/ω is the fractional rate of slowing of rotation, c is a mean gas molecular velocity, α is the coefficient of expansion of the ball, and T is the ball's temperature.

10.10.4 Membrane Vacuum Sensors

In spite that a membrane was considered not flexible enough for measuring high degree vacuum, recent developments in MEMS technologies allowed fabrication of

silicone membranes that allow a full-scale pressure range from 0 to 0.5 Torr with a sub-μTorr resolution [21]. The membrane is used as a plate in a capacitive sensor, similar to those described in this chapter (Sect. 10.6).

References

1. Benedict RP (1984) Fundamentals of temperature, pressure, and flow measurements, 3rd edn. Wiley, New York
2. Plandtl L (1952) Essentials of fluid dynamics. Hafner, New York
3. Neubert HKP (1975) Instrument transducers. An introduction to their performance and design, 2nd edn. Clarendon, Oxford
4. Clark SK, Wise KD (1979) Pressure sensitivity in anisotropically etched thin-diaphragm pressure sensor. IEEE Trans Electron Devices ED-26:1887–1896
5. Tufte ON, Chapman PW, Long D (1962) Silicon diffused-element piezoresistive diaphragm. J Appl Phys 33:3322–3327
6. Kurtz AD, Gravel CL (1967) Semiconductor transducers using transverse and shear piezoresistance. In: Proc. 22nd ISA conf., No. P4-1 PHYMMID-67, Sept.
7. Tanigawa H, Ishihara T, Hirata M, Suzuki K (1985) MOS integrated silicon pressure sensor. IEEE Trans Electron Devices ED-32(7):1191–1195
8. Petersen K, Barth P, Poydock J, Brown J, Mallon J Jr, Bryzek J (1988) Silicon fusion bonding for pressure sensors. In: Rec. proc. IEEE solid-state sensor actuator workshop, pp 144–147
9. Proud R (1991) VRP transducers for low-pressure measurement. Sensors, Feb 1991, pp 20–22
10. Wolthuis RA, Mitchell GL, Saaski E, Hratl JC, Afromowitz MA (1991) Development of medical pressure and temperature sensors employing optical spectral modulation. IEEE Trans Biomed Eng 38(10):974–981
11. Hälg B (1991) A silicon pressure sensor with an interferometric optical readout. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers, pp 682–684, IEEE, 1991
12. Vaughan JM (1989) The Fabry–Perot interferometers. Bristol, Adam Hilger
13. Saaski EW, Hartl JC, Mitchell GL (1989) A fiber optic sensing system based on spectral modulation. Paper #86-2803, ©ISA
14. Von Smoluchovski M (1911) Ann Physik 35:983
15. Fraden J (2009) Detector of low levels of gas pressure and flow. US Patent 7,490,512, 17 Feb
16. Kermode AC (2006) Mechanics of flight. In: Barnard RH, Philpott DR (eds) 11th edn. Prentice Hall, Harlow
17. Buckley OE (1916) Proc Natl Acad Sci USA 2:683
18. Leck JH (1957) Pressure measurement in vacuum systems. Chapman & Hall, London, pp 70–74
19. Bayard RT, Alpert D (1950) Rev Sci Instrum 21:571
20. Fremery JK (1946) Vacuum 32:685
21. Zhang Y et al (2001) An ultra-sensitive, high-vacuum absolute capacitive pressure sensor. In: 14th IEEE international conference on micro electro mechanical systems (Cat. No. 01CH37090), Technical Digest, pp 166–169.
22. Goehner R, Drubetsky E, Brady HM, Bayles WH Jr (2000) Vacuum measurement. In: Webster J (ed) Mechanical variables measurement. CRC Press LLC, Boca Raton, FL
23. Di Giovanni M (1982) Flat and corrugated diaphragm design handbook. Marcel Dekker, Inc., New York.

Chapter 11
Flow Sensors

A distinguished scientist was asked which two questions he would ask God?
– Oh, I would ask him to explain the theory which links quantum mechanics and general relativity.
– And the second question? Would you ask Him to explain turbulence?
– No, I don't wish to embarrass Him...

11.1 Basics of Flow Dynamics

One of the fundamentals of physics is that mass is a conserved quantity. It cannot be created or destroyed. In the absence of sources or sinks of mass, its quantity remains constant regardless of boundaries. However, if there is influx or outflow of mass through the boundaries, the sum of influx and efflux must be zero. Whatever mass comes in, must go out. When both are measured over the same interval of time, mass entering the system (M_{in}) is equal to mass leaving the system (M_{out}) [1]. Therefore,

$$\frac{dM_{in}}{dt} = \frac{dM_{out}}{dt}. \tag{11.1}$$

In mechanical engineering, moving media whose flow is measured are liquids (water, oil, solvents, gasoline, etc.), air, gases (oxygen, nitrogen, CO, CO_2, methane CH_4, water vapor, etc.).

In a steady flow, the velocity at a given point is constant in time. We can draw a streamline through every point in a moving medium (Fig. 11.1a). In steady flow, the line distribution is time-independent. A velocity vector is tangent to a stream line in every point z. Any boundaries of flow which envelop a bundle of streamlines are called a tube of flow. Since the boundary of such a tube consists of streamlines, no fluid (gas) can cross the boundary of a tube of flow and the tube behaves something

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_11, © Springer Science+Business Media, LLC 2010

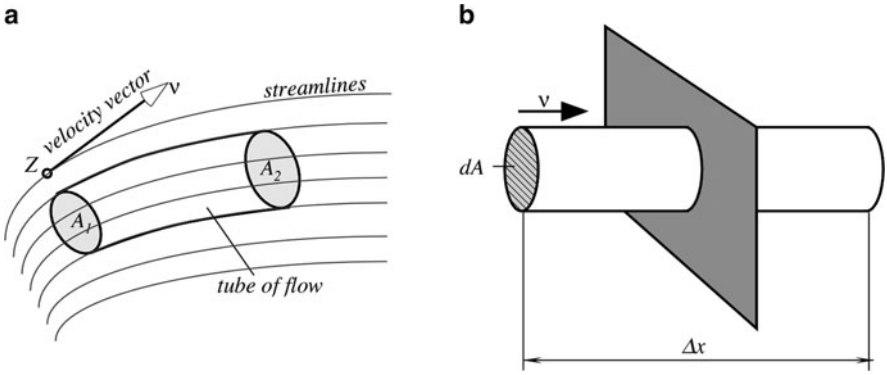

Fig. 11.1 Tube of flow (**a**) and flow of a medium through a plane (**b**)

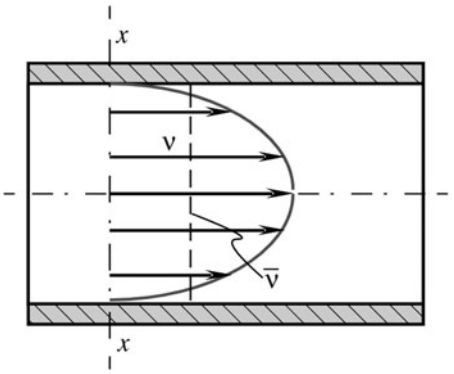

Fig. 11.2 Profile of velocity of flow in a pipe

like a pipe of some shape. The flowing medium can enter such a pipe at one end, having cross-section A_1 and exit at the other through cross-section A_2. The velocity of a moving material inside a tube of flow will in general have different magnitudes at different points along the tube.

The volume of moving medium passing a given plane (Fig. 11.1b) in a specified time interval Δt is

$$\Lambda = \frac{V}{\Delta t} = \int \frac{\Delta x}{\Delta t} dA = \int v dA \tag{11.2}$$

where v is the velocity of moving medium which must be integrated over area A. while Δx is the displacement of volume V. Figure 11.2 shows that the velocity of liquid or gas in a pipe may vary over the cross-section. It is often convenient to define an average velocity

$$v_a = \frac{\int v dA}{A} \tag{11.3}$$

When measuring the velocity by a sensor whose dimensions are substantially smaller than the pipe size, one should be aware of a possibility of erroneous detection of either too low or too high velocity, while the average velocity, v_a, is somewhere in-between. A product of the average velocity and a cross-sectional area is called flux or flow rate. Its SI unit is m^3/s. The U.S. Customary System unit is ft^3/s. The flux can be found by rearranging (11.3)

$$Av_a = \int v dA \tag{11.4}$$

What a flow sensor usually measures is v_a. Thus, to determine a flow rate, a cross-section area of tube of flow A must be known, otherwise the measurement is meaningless.

The measurement of flow is rarely conducted for the determination of a displacement of volume. Usually, what is needed is to determine the flow of mass rather than volume. Of course, when dealing with virtually incompressible fluids (water, oil, etc.), either volume or mass can be used. A relationship between mass and volume for an incompressible material is through density ρ

$$M = \rho V. \tag{11.5}$$

The densities of some materials are given in Table A.12. The rate of mass flow is defined as

$$\frac{dM}{dt} = \rho A \overline{v} \tag{11.6}$$

The SI unit for mass flow is kg/s while the U.S. Customary System unit is lb/s. For a compressible medium (gas) either mass flow or volume flow at a given pressure should be specified.

There is a great variety of sensors which can measure flow velocity by determining the rate of displacement either mass, or volume. Whatever sensor is used, inherent difficulties of the measurement make the process a complicated procedure. It is necessary to take into consideration many of the natural characteristics of the medium, its surroundings, barrel and pipe shapes and materials, medium temperature and pressure, etc. When selecting any particular sensor for the flow measurement, it is advisable to consult with the manufacturer's specifications and very carefully considering the application recommendations for a particular sensor. In this book, we do not cover such traditional flow measurement systems as a moving vane or turbine type meters. It is of interest to us to consider sensors without moving components that introduce either no or little restriction into the flow.

11.2 Pressure Gradient Technique

A fundamental equation in fluid mechanics is Bernoulli equation,[1] which is strictly applicable only to steady flow of nonviscous, incompressible medium

$$p + \rho\left(\frac{1}{2}v_a^2 + gy\right) = const, \tag{11.7}$$

where p is the pressure in a tube of flow, $g = 9.80665$ m/s^2 $= 32.174$ ft/s^2 is the gravity constant, and y is the height of medium displacement. Bernoulli's equation allows us to find fluid velocity by measuring pressures along the flow.

The pressure gradient technique (of flow measurement) essentially requires an introduction a flow resistance. Measuring the pressure gradient across a known resistor allows to calculate a flow rate. The concept is analogous to Ohm's law: voltage (pressure) across a fixed resistor is proportional to current (flow). In practice, the restricting elements that cause flow resistances are orifices, porous plugs, and Venturi tubes (tapered profile pipes). Figure 11.3 shows two types of flow resistors. In the first case it is a narrow in the channel, while in the other case, there is a porous plug, which somewhat restricts the medium flow. A differential pressure sensor is positioned across the resistor. When a moving mass enters the higher resistance area, its velocity increases in proportion to the resistance increase:

$$v_{1a}{=}v_{2a}R. \tag{11.8}$$

The Bernoulli equation defines differential pressure as[2]

$$\Delta p = p_1 - p_2 = \frac{\rho}{2}\left(v_{2a}^2 - v_{1a}^2\right) = k\frac{\rho}{2}v_{2a}^2(1 - R^2) \tag{11.9}$$

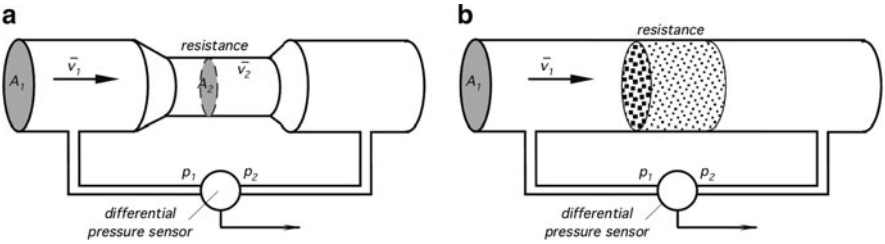

Fig. 11.3 Two types of flow resistors: a narrow channel (**a**) and a porous plug (**b**)

[1]The Bernoulli's principle is named after the Dutch-Swiss mathematician Daniel Bernoulli who published his principle in his book *Hydrodynamica* in 1738.

[2]It is assumed that both pressure measurements are made at the same height ($y = 0$), which is usually the case.

where k is the correction coefficient, which is required because the actual pressure p_2 is slightly lower than the theoretically calculated. From (11.9) the average velocity can be calculated as

$$v_{2a} = \frac{1}{\sqrt{k(1-R^2)}}\sqrt{\frac{2}{\rho}\Delta p} \tag{11.10}$$

To determine the mass flow rate per unit time, for an incompressible medium, (11.10) is simplified to

$$q = \xi A_2 \sqrt{\Delta p} \tag{11.11}$$

where ξ is a scaling coefficient, which is determined through calibration. The calibration must be done with a specified liquid or gas over an entire operating temperature range since the value of ξ may be different at different temperatures. It follows from the above that the pressure gradient technique essentially requires the use of either one differential pressure sensor or two absolute sensors. If a linear representation of the output signal is required, a square root-extraction must be used. The root-extraction can be performed in a microprocessor by using one of the conventional computation techniques. An advantage of the pressure gradient method is in the absence of moving components and use of standard pressure sensors, which are readily available. A disadvantage is in the restriction of flow by resistive devices.

A microflow sensor can be constructed by utilizing a capacitive pressure sensor [2] shown in Fig. 11.4. An operating principle of the sensor is based on a pressure gradient technique. The sensor was fabricated using silicon micromachining and defused boron etch-stops to define the structure. The gas enters the sensor's housing

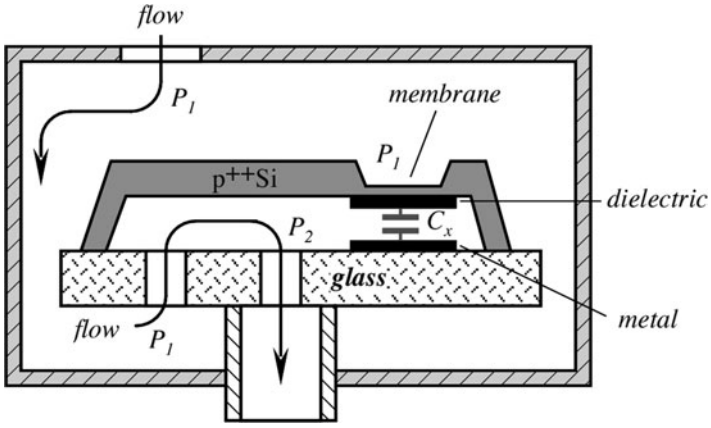

Fig. 11.4 Structure of a gas microflow sensor utilizing capacitive pressure sensor (adapted from [2])

at pressure P_1 through the inlet, and the same pressure is established all around the silicon plate, including the outer side of the etched membrane. The gas flows into the microsensor's cavity through a narrow channel having a relatively high-pressure resistance. As a result, pressure P_2 inside the cavity is lower than P_1, thus creating a pressure differential across the membrane. Therefore, the flow rate can be calculated from (11.10).

The pressure differential causes the membrane deflection that is measured by a capacitive pressure sensor. The capacitance C_x is formed of a thin, stress compensated, p^{2+} boron-doped silicon membrane suspended above a metal plate. Pressure differential changes capacitance C_x between the metal plate and the silicon structure with a resolution of 1 mTorr/1 fF with a full pressure of about 4 Torr. An overall resolution of the sensor is near 14–15 bits and the accuracy of pressure measurement about 9–10 bits. At approximately twice the full-scale pressure differential, the membrane touches the metal plate, hence a dielectric layer is required to prevent an electric short, while the substrate glass plate protects the membrane from rupturing. A capacitance measurement circuit (see Fig. 5.32) is integrated with the silicon plate using a standard CMOS technology.

11.3 Thermal Transport Sensors

A good method of measuring a flow would be by somehow marking the flowing medium and detecting the movement of the mark. For example, a mark can be a floating object, which can move with the medium while being stationary with respect to the medium. The time which it would take the object to move with the flow from one position to another could be used for the calculation of the flow rate. Such an object may be a float, radioactive element, or die, such as colored fluid (color smoke or liquid paint). Also, the mark can be a different gas or liquid whose concentration and rate of dilution can be detectable by appropriate sensors.

In medicine, a die-dilution method of flow measurement is used for studies in hemodynamics. In most instances, however, placing any foreign material into the flowing medium is either impractical or forbidden for some other reasons. An alternative would be to change certain physical properties of the moving medium and to detect the rate of displacement of the changed portion or rate of its dilution. Usually, the best physical property, which can be easily modified without causing undesirable effects, is temperature.

The sensors that detect the rate of heat dissipation in flowing media are called thermal transport flowmeters or *thermoanemometers*. Thermal transport flowmeters are far more sensitive than other types and have a broad dynamic range. They can be employed to measure very minute gas or liquid displacements as well as fast and strong currents. The major advantages of these sensors are the absence of moving components and an ability to measure very low flow rates. "Paddle wheel," hinged vane, and pressure differential sensors have low and inaccurate outputs at low flow rates. If a small diameter of tubing is required, as in automotive, aeronautic,

medical, and biological applications, sensors with moving components become mechanically impractical. In these applications, thermal transport sensors are indispensable. Another advantage of these sensors is their usefulness for detecting the material change in composition because they are sensitive to heat transport in a media that is typically altered by a changing composition or chemical reaction.

A thermoanemometer design determines its operating limits. At a certain velocity, the molecules of a moving medium while passing near a heater do not have sufficient time to absorb enough thermal energy for developing a temperature differential between two detectors. The upper operating limits for the thermal transport sensors usually are determined experimentally. For instance, under normal atmospheric pressure and room temperature (about 20°C), the maximum air velocity which can be detected by a thermal transport sensor is in the range of 60 m/s (200 ft/s).

The pressure and temperature of a moving medium, especially of gases, make a strong contribution to the accuracy of a volume rate calculation. It is interesting to note that for the mass flow meters, pressure makes very little effect on the measurement as the increase in pressure results in a proportional increase in mass.

A data processing system for the thermal transport sensing must receive at least three variable input signals: a flowing medium temperature, a temperature differential, and a heating power signal. These signals are multiplexed, converted into digital form and processed by a computer to calculate characteristics of flow. Data are usually displayed in velocities (m/s or ft/s), volume rates (m^3/s or ft^3/s), or mass rate (kg/s or lb/s).

11.3.1 Hot-Wire Anemometers

The oldest and best-known thermal transport flow sensors are the hot-wire and later developed hot-film anemometers [3]. They have been used quite extensively for measurements of turbulence levels in wind tunnels, flow patterns around models, and blade wakes in radial compressors. A hot-wire thermoanemometer is a single-part sensor as opposed to two- and three-part sensors as described below. The key element of this sensor is a heated wire having typical dimensions 0.00015–0.0002 in. (0.0038–0.005 mm) in diameter and 0.040–0.080 in. (1.0–2.0 mm) in length. The wire resistance typically is between 2 and 3 Ω. The operating principle is based on warming up the wire by electric current to 200–300°C, well above the flowing media temperature and then measuring temperature of the wire. Such a high temperature, that typically is well over temperatures of the flowing media, makes the sensor little sensitive to the media temperature and thus no media temperature compensation is required. Under the no-flow condition, temperature of the wire will be constant, but when the media flows the wire will be cooled. The stronger the flow the stronger is the cooling. The advantage of the hot wire and hot film probes is in their fast speed responses; they can resolve frequencies up to 500 Hz.

There are two types of methods to control temperature and measuring the cooling effect: a constant voltage and a constant temperature. In the former method, reduction in the wire temperature is measured, while in the latter case, the temperature is maintained constant at any reasonable flow rate by the increase in supplied electric power. That power is the measure of the flow rate. In a hot-wire anemometer, the wire has a positive temperature coefficient and thus is used for a dual purpose: to elevate temperature above the media temperature (so it will be a cooling effect) and also to measure that temperature because the wire resistance goes down when the wire cools. Figure 11.5 shows a simplified bridge circuit for the constant temperature method. This is a null-balancing bridge that is based on the principle described in Sect. 5.10.3 (Fig. 5.38b).

The feedback from a servo amplifier keeps the bridge in a balanced state. Resistors R_1–R_3 are constant, while R_w represents resistance of the hot wire and is temperature dependent. Drop in the wire temperature t_w causes temporary drop in R_w and a subsequent reduction in the bridge voltage $-e$ that is applied to the negative input of the servo amplifier. This leads to increase in V_{out}, which is applied to the bridge as a feedback. When V_{out} goes up, current i through the wire increases, leading to increase in temperature. This restores the wire temperature when flowing media attempt to cool it, so t_w remains constant over the entire flow rate range. The feedback voltage V_{out} is the output signal of the circuit and the measure of the mass flow rate. The faster the flow, the higher the voltage.

Under a steady flow rate, the electric power Q_e supplied to the wire is balanced by the out-flowing thermal power Q_T carried by the flowing media due to a convective heat transfer. That is,

$$Q_e = Q_T. \tag{11.12}$$

Considering the heating current i, the wire temperature t_w, temperature of the fluid t_f, the wire surface area A_w, and the heat transfer coefficient h, we can write the balance equation

$$i^2 R_w = hA_w\left(t_w - t_f\right). \tag{11.13}$$

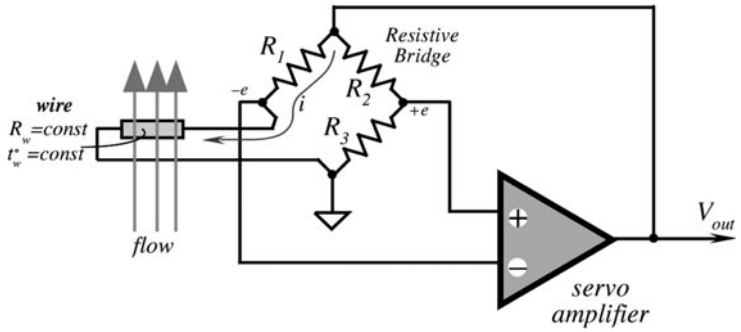

Fig. 11.5 Null-balanced bridge for a constant temperature hot-wire anemometer

In 1914 King [4] developed a solution of a heat loss from an infinite cylindrical body in an incompressible low Reynolds number flow that may be written as

$$h = a + bv_f^c, \tag{11.14}$$

where a and b are constant and $c \approx 0.5$. This equation is known as King's law.

Combining the above three equations allows us to eliminate the heat transfer coefficient h:

$$a + bv_f^c = \frac{i^2 R_w}{A_w\left(t_w - t_f\right)}. \tag{11.15}$$

Considering that $V_{out} = i(R_w + R_1)$ and $c = 0.5$, we can solve this equation for the output voltage as function of the fluid velocity:

$$V_{out} = (R_w + R_1)\sqrt{\frac{A_w(a + b\sqrt{v})(t_w - t_f)}{R_w}}. \tag{11.16}$$

Note that thanks to high temperature gradient $(t_w - t_f)$, the output signal depends little of the media temperature t_f. For efficient operation, the temperature gradient $(t_w - t_f)$ and the sensor surface area should be as large as practical.

Since the King's law was derived for an infinite cylinder, its applicability to a practical sensor should be taken with a grain of salt. The hot wire is relatively short (no more than 2 mm) and must be somehow supported by the probe and held steady while inside the flow. Also, electrical resistance of the wire should be relatively low to allow for heating by electric current. At the same time, the wire should have as large thermal coefficient as possible. A very careful design is required to meet these requirements. Heat can be lost from the wire not only by means of convection (a useful effect) but also via thermal radiation and thermal conduction (interfering effects). While thermal radiation is usually negligibly small and can be ignored, conductive heat loss via the support structure may be comparable or even greater than the convective loss. Thus, the heated wire must have as large thermal resistance to the support structure as physically possible. This poses series challenges to the sensor designer.

A typical design of the hot-wire sensor is shown in Fig. 11.6a. The most common wire materials are tungsten, platinum, and a platinum–iridium alloy. Tungsten wires are strong and have a high temperature coefficient of electrical resistance (0.004/°C). However, they cannot be used at high temperatures in many gases because of poor oxidation resistance. Platinum has good oxidation resistance, has a good temperature coefficient (0.003/°C), but is very weak, particularly at high temperatures. The platinum–iridium wire is a compromise between tungsten and platinum with good oxidation resistance, and more strength than platinum, but it has a low temperature coefficient of electrical resistance (0.00085/°C). Tungsten is presently the more popular hot wire material. A thin platinum coating is usually applied to improve bond with the plated ends and the support needles. The needles

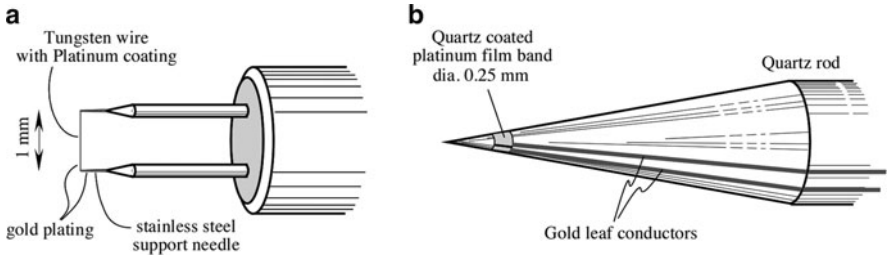

Fig. 11.6 Hot-wire probe (**a**) and a conical hot-film probe (**b**)

should be thin but strong and have high thermal resistance (low thermal conductivity) to the probe body. Stainless steel is the most often used material. Hot-wires probes are expensive, extremely fragile, and can be damaged easily mechanically or by an excessive electric pulse.

A hot-film sensor is essentially a conducting film deposited on an insulator, such as a ceramic substrate. The sensor shown in Fig. 11.6b is a quartz cone with a platinum film on the surface. Gold plating on the sides of the cone provides electrical connection. When compared with hot wires, the hot-film sensor has the following advantages:

- Better frequency response (when electronically controlled) than a hot wire of the same diameter because the sensitive part of the film sensor has a larger surface area
- Lower heat conduction to the supports for a given length to diameter ratio due to the low thermal conductivity of the substrate material. A shorter sensing length can thus be used
- More flexibility in sensor configuration. Wedge, conical, parabolic, and flat surface shapes are available
- Less susceptible to fouling and easier to clean. A thin quartz coating on the surface resists accumulation of foreign material

The metal film thickness on a typical film sensor is less than 1,000 Å, thus the physical strength and the effective thermal conductivity is determined almost entirely by the substrate material. Most films are made of platinum due to its good oxidation resistance and the resulting long-term stability. A better ruggedness and stability of film sensors have led to their use for many measurements that have previously been very difficult with the more fragile and less stable hot wires. The hot film probes have been made on cones, cylinders, wedges, parabolas, hemispheres, and flat surfaces. Cylindrical film sensors that are cantilever-mounted are also made. This is done by making the cylindrical film sensor from a quartz tube and running one of the electrical leads through the inside of the tube. The cone-shaped sensor of Fig. 11.6b is used primarily in water applications where its shape is particularly valuable in preventing lint and other fibrous impurities from getting entangled with sensor. The cone can be used in relatively contaminated water, while cylindrical sensors are more applicable when the water has been filtered.

11.3.2 Three-Part Thermoanemometer

The thermal anemometer shown in Fig. 11.7 is used primarily for liquids but also may be useful for gases. It is a very rugged and contamination-resistant device. This sensor is comprised of three small tubes immersed into a moving medium. Two tubes contain temperature detectors R_o and R_s. The detectors are thermally coupled to the medium and are thermally isolated from the structural elements and the pipe where the flow is measured. In between two detectors, a heating element is positioned. Both detectors are connected to electrical wires through tiny conductors to minimize thermal loss through conduction (Fig. 11.7b).

The sensor operates as follows. The first temperature detector R_o measures the temperature of the flowing medium. Downstream from the first sensor, the heater warms up the medium and the elevated temperature is measured by the second temperature detector R_s. In a still medium, heat would be dissipated from the heater through media to both detectors. In a medium with a zero flow, heat moves out from the heater mainly by thermal conduction and gravitational convection. Since the heater is positioned closer to the R_s detector, that detector will register higher temperature. When the medium flows, heat dissipation increases due to forced convection. The higher the rate of flow the higher the heat dissipation and the lower temperature will be registered by the R_s detector. Heat loss is measured and converted into the flow rate of medium.

A fundamental relationship of the thermoanemometry is based on King's law as described above for the hot-wire anemometer. An incremental heat change is

$$\Delta Q = kl\left(1 + \sqrt{\frac{2\pi\rho cdv}{k}}\right)\left(t_s - t_f\right), \tag{11.17}$$

where k and c are the thermal conductivity and specific heat of a medium at a given pressure, ρ is the density of the medium, l and d are the length and diameter of the second temperature sensor, t_s is the surface temperature of the second sensor, t_f is

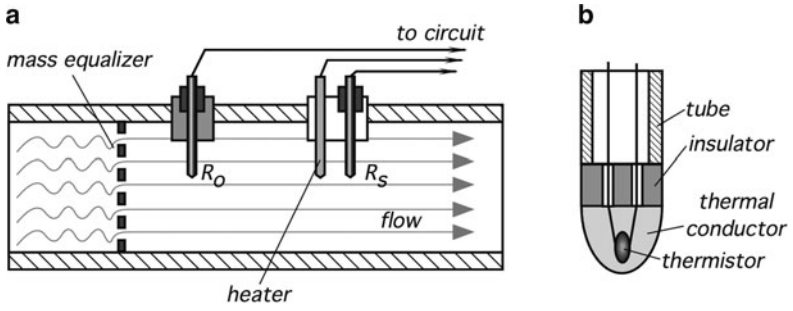

Fig. 11.7 Three-part thermoanemometer.
Basic two-sensor design (**a**); cross-sectional view of a temperature sensor (**b**)

the temperature of the first sensor (media temperature), and v is the velocity of the medium. Collis and Williams experimentally proved [5] that King's theoretical law needs some correction. For a cylindrical sensor with $l/d \gg 1$, a modified King's equation yields the velocity of the medium:

$$v = \frac{K}{\rho}\left(\frac{dQ}{dt}\frac{1}{t_s - t_f}\right)^{1.87}, \tag{11.18}$$

where K is the calibration constant. It follows from the above that to measure a flow, a temperature gradient between the second sensor and the moving medium, and dissipated heat must be measured. Then velocity of the fluid or gas becomes although nonlinear, but a quite definitive function of thermal loss (Fig. 11.8a).

For accurate temperature measurements in a flowmeter, any type of temperature detector can be used: resistive, semiconductor, optical, etc. (Chap. 17). Nowadays, however, the majority of manufacturers use resistive sensors. In industry and scientific measurements, RTDs are the prime choice as they assure higher linearity, predictable response, and long-term stability over broader temperature range. In medicine, thermistors are often preferred thanks to their higher sensitivity. Whenever a resistive temperature sensor is employed, especially for a remote sensing, a four-wire measurement technique should be seriously considered. The technique is a solution for a problem arising from a finite resistance of connecting wires, which may be a substantial source of error, especially with low resistance temperature sensors like RTDs. See Sect. 5.12.2 for the description of a four-wire method.

While designing thermal flow sensors, it is important to assure that the medium moves through the detectors without turbulence in a laminar well mixed flow. The sensor is often supplied with mixing grids or turbulence breakers which sometimes are called mass equalizers (Fig. 11.7a).

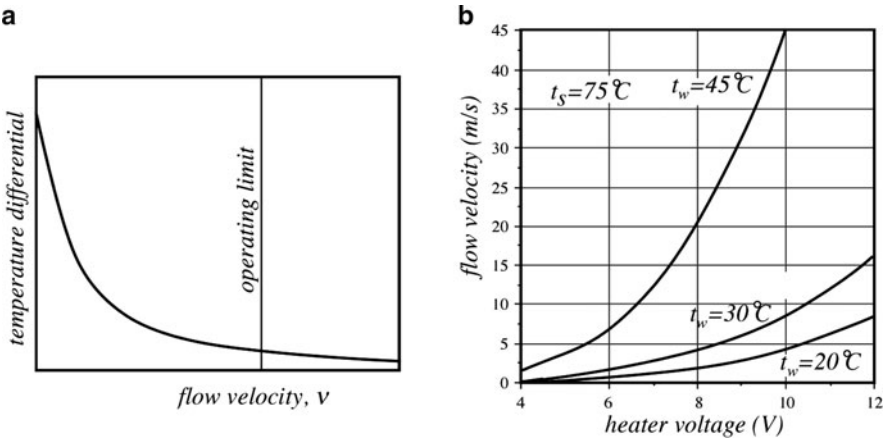

Fig. 11.8 Transfer function of a thermoanemometer (**a**) and calibration curves for a self-heating sensor in a thermoanemometer (**b**) for three different levels of heat

11.3.3 Two-Part Thermoanemometer

The hot-wire and hot-film anemometers that were described above are the fast-response sensors. However, they are the extremely delicate and expensive devices that are sensitive to many airborne contaminants like dust and smoke. In many applications, gas flow should be monitored continuously for a long time without the need for a fast response. Such sensors should be more resistant to gas impurities and mechanically rugged. Since the speed of the sensor response can be sacrificed for the sake of robustness, a different design approach can be taken. Yet, the key requirement of maximizing a thermal resistance between a heated element and the support structure of the probe should be maintained, as in all thermal transport flow sensors. Three functions should be accomplished by the sensor: measuring temperature of the flowing media, heating the media, and monitoring the cooling effect caused by the flow. Figure 11.9a shows a two-part thermoanemometer [6, 7] where one part is the media temperature reference sensor S_1, wile the other part is a combination of the heating H and temperature sensing S_2 elements that are kept in an intimate thermal coupling with each other. In other words, the second temperature sensor measures temperature of the heater.

Both temperature sensors are the thick-film NTC thermistors (see Sect. 16.3.3.1) printed on the ceramic substrates. These substrates form two sensing "fingers" that are visible in Fig. 11.9a. The second substrate also has a heating resistor $H = 150\ \Omega$ printed over the thermistor layer S_2 with an electrical isolating barrier in-between. Both fingers are coated with thing layers of protective glass or thermally conductive epoxy. The thermistors are connected into a Wheatstone bridge (Fig. 11.10a) where two fixed resistors R_3 and R_4 comprise the other two arms of the bridge. To generate heat, the following condition must be met: $R_4 < R_3$. The warm finger is thermally decoupled from the reference finger and both fingers are subjected to air flow. This sensor is not very fast. It can be characterized by a relatively long time constant of about 0.5 s, yet for many applications this is sufficiently short.

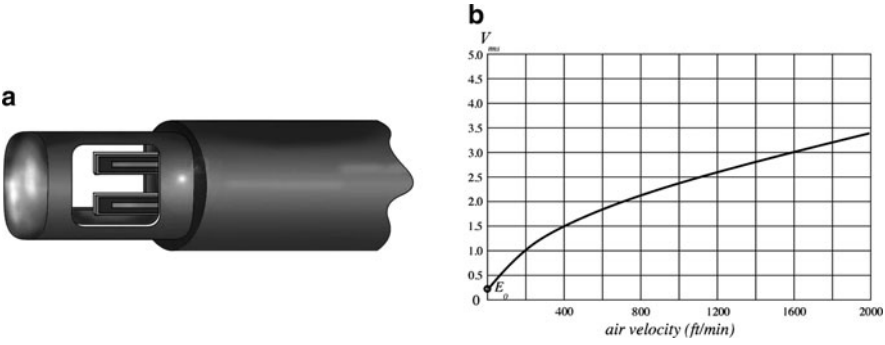

Fig. 11.9 Two-part thermoanemometer probe (**a**) and transfer function (**b**) (Courtesy of Clean-Alert, LLC, Oberlin, OH)

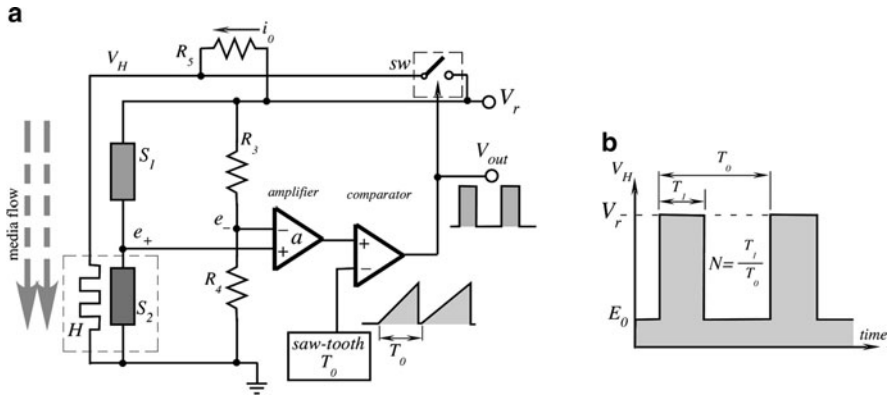

Fig. 11.10 Control circuit of a thermoanemometer with PWM modulator (**a**); transfer function for the r.m.s. of PWM signal (**b**)

The thermistor S_2 is warmed up above the media (air) temperature by a constant increment $\Delta t = 5$–$10°$C. During operation, air moves across both thermistors S_1 and S_2 and removes thermal energy from the warmer thermistor S_2 in relation to the airflow velocity. The convective cooling is compensated by electric power provided to the heater H from a feedback circuit that is similar to Fig. 11.5. However there is a difference; the heating voltage V_H is a pulse-width-modulated (PWM) feedback to the heater whose r.m.s. value provides a Joule heat. The PWM signal is formed with the help of a saw-tooth generator having period T_0 of about 4 ms. The pulses from the comparator control the switch sw that connects the heater to reference voltage V_r. The wider the PWM pulses, the more heat is generated. This PWM signal is the output of the sensor where the duty cycle N represents the air flow rate.

The sensor is quite rugged and both sensing fingers are securely supported by the interior of the probe. The price for that is a not too high thermal resistance r between the finger and probe body. As a result, a significant thermal power P_L is lost to the supporting structure via thermal conduction. To compensate for this undesirable effect, resistor R_5 provides additional current i_0 to H when the switch sw is open. This is visible in Fig. 11.10b as a voltage pedestal E_0 across H.

Initially, at the moment when the reference voltage V_r is just being turned on, the thermistors are at the same air temperature and thus have nearly equal resistances. The bridge is disbalanced because $R_4 < R_3$. The bridge differential voltage $(e_+ - e_-)$ is amplified with gain a, compared with the saw-tooth signal and controls the gating switch sw, resulting in the output voltage pulses V_H, which are applied to the heater H. This develops a Joule heat in the heater H and warms up the thermistor S_2. Temperature of S_2 goes up by Δt above the air temperature and resistance of S_2 drops till the moment when the bridge enters a balanced state:

$$\frac{S_1}{S_2} = \frac{R_3}{R_4}. \tag{11.19}$$

The balance is kept by the feedback circuit as long as there is no change in the air movement near the sensing thermistors so the ratio of (11.19) is satisfied. Change in airflow rate (change in cooling) disbalance the bridge and subsequently modulates the duty cycle N of the PWM signal to restore the ratio of (11.19). Thus, value of N reflects the airflow rate.

To obtain the sensor's transfer function, remember that the heater H and thermistor S_2 lose thermal energy to the probe by means of thermal conduction at a rate

$$P_L = \frac{\Delta t}{r}, \tag{11.20}$$

where r is a thermal resistance (°C/W) to the support structure. A typical value of r is on the order of 50°C/W and the goal is to make that number as large as possible. The conductive loss compensating power to H via R_5 when sw is open is

$$P_0 = \frac{E_0^2}{H}(1-N)^2 \tag{11.21}$$

The flowing air results in a convective heat loss from S_2 that is defined as

$$P_a = k\nu\Delta t, \tag{11.22}$$

where v is the air velocity and k is the scaling factor. To compensate for convective cooling, Joule heat power is delivered to H from the PWM feedback circuit:

$$P_f = \frac{N^2 V_r^2}{H}. \tag{11.23}$$

The law of conservation of energy demands that under a steady-state condition

$$P_L + P_a = P_0 P_f. \tag{11.24}$$

Substituting (11.20–11.23) into (11.24) we arrive at

$$\frac{\Delta t}{r} + k\nu\Delta t = i_0^2 H(1-N)^2 + \frac{N^2 V_r^2}{H} = \frac{E_0^2}{H}(1-N)^2 + \frac{N^2 V_r^2}{H}, \tag{11.25}$$

from which we derive the output value of the PWM duty cycle N:

$$N = \sqrt{\frac{\left(\frac{\Delta t}{r} + k\nu\Delta t\right)H - i_0^2 H^2}{V_r^2 - i_0^2 H^2}} \approx \frac{\sqrt{\left(\left(\frac{1}{r} + k\nu\right)\Delta t - i_0^2 H\right)H}}{V_r}. \tag{11.26}$$

Since the value in parenthesis is always positive the following condition must be met:

$$\frac{\Delta t}{r} \geq i_0^2 H = \frac{V_r^2 H}{(H + R_5)^2} \tag{11.27}$$

The sensor's response is shown in Fig. 11.9b. In a practical design, thermistors S_1 and S_2 may have manufacturer's tolerances that should be compensated for by trimming one of the resistors R_3, R_4.

11.3.4 Microflow Thermal Transport Sensors

In some applications, such as process control in precise semiconductor manufacturing, chemical and pharmaceutical industries, and biomedical engineering, and miniaturized gas flow sensors are encountered with an increasing frequency. Most of them operate on the method of a thermal transport and are fabricated from a silicon crystal by using micromachining technology (MEMS). Many of the microflow sensors use thermopiles as temperature sensors [8, 9].

A cantilever design of a microflow sensor is shown in Fig. 11.11. Thickness of the cantilever may be as low as 2 μm. It is fabricated in the form of a sandwich consisting of layers of field oxide, CVD oxide, and nitrate [10]. The cantilever sensor is heated by an imbedded resistor with a rate of 26 K/mW of applied electric power, and a typical transfer function of the flow sensor has a negative slope of about 4 mV(/m/s).

The heat is removed from the sensor by three means: conductance L_b through the cantilever beam, gas flow $h(v)$ and thermal radiation, which is governed by the Stefan–Boltzmann law:

$$P = L_b(T_s - T_b) + h(v)(T_s - T_\mathrm{b}) + a\sigma\varepsilon(T_s^4 - T_b^4), \tag{11.28}$$

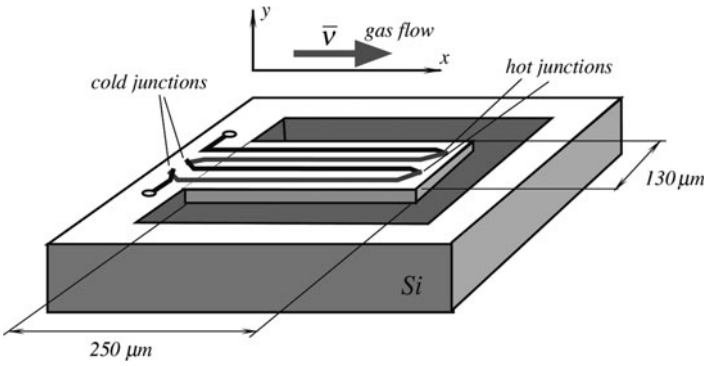

Fig. 11.11 Micromachined gas flow sensor

where σ is the Stefan–Boltzmann constant, a is the area along which the beam-to-gas heat transfer occurs, ε is surface emissivity, and v is the gas velocity. From the principles of energy and particle conservation we deduce a generalized heat transport equation governing the temperature distribution $T(x, y)$ in the gas flowing near the sensor's surface

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = \frac{vnc_p}{k_g}\frac{\partial T}{\partial x} \quad \text{for } y > 0, \tag{11.29}$$

where n is the gas density, c_p is the molecular gas capacity, and k_g is the thermal conductivity of gas. It can be shown that solution of this equation for the boundary condition of vanishing thermal gradient far off the surface is [10]

$$\Delta V = B\left(\frac{1}{\sqrt{\mu^2 + 1}} - 1\right), \tag{11.30}$$

where V is the input voltage, B is a constant, and $\mu = Lvnc_p/2\pi k_g$, and L is the gas-sensor contact length. This solution coincides very well with the experimental data.

Another design of a thermal transport microsensor is shown in Fig. 11.12a [11] where the titanium films having thickness of 0.1 μm serve as both the temperature sensors and heaters. The films are sandwiched between two layers of SiO_2. Titanium was used because of its high temperature coefficient of resistance (TCR) and excellent adhesion to SiO_2. Two microheaters are suspended with four silicon girders at a distance of 20 μm from one another. The Ti film resistance is about 2 kΩ. Figure 11.12b shows a simplified interface circuit diagram for the sensor, which exhibits an almost linear relationship between the flow and output voltage change ΔV.

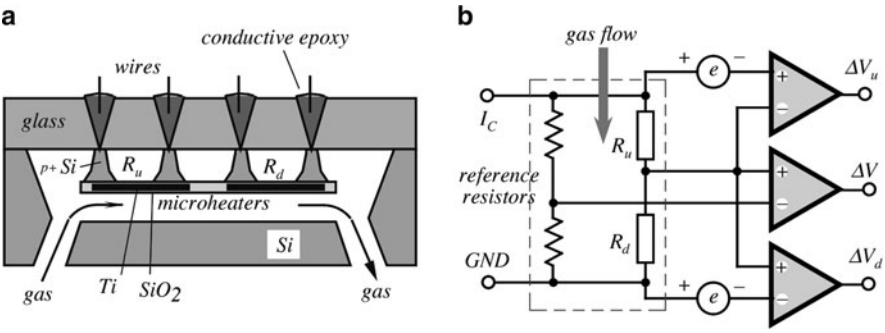

Fig. 11.12 Gas microflow sensor with self-heating titanium resistors sensor design (**a**); interface circuit (**b**). R_u and R_d are resistances of the up- and downstream heaters respectively (adapted from [10])

11.4 Ultrasonic Sensors

Flow can be measured by employing ultrasonic waves. The main idea behind the principle is the detection of frequency or phase shift caused by flowing medium. One possible implementation is based on the Doppler effect (see Sect. 6.2 for the description of the Doppler effect), while the other relies on the detection of the increase or decrease in effective ultrasound velocity in the medium. Effective velocity of sound in a moving medium is equal to the velocity of sound relative to the medium plus the velocity of the medium with respect to the source of the sound. Thus, a sound wave propagating upstream will have a smaller effective velocity, while the sound propagating downstream will have a higher effective velocity. Since the difference between the two velocities is exactly twice the velocity of the medium, measuring the upstream–downstream velocity difference allows us to determine the velocity of the flow.

Figure 11.13a shows two ultrasonic generators positioned at opposite sides of a tube of flow. Piezoelectric crystals are usually employed for that purpose. Each crystal can be used for either the generation of the ultrasonic waves (motor mode), or for receiving the ultrasonic waves (generator mode). In other words, the same crystal, when needed, acts as a "speaker" or a "microphone."

Two crystals are separated by distance D and positioned at angle Θ with respect to flow. Also, it is possible to place small crystals right inside the tube along the flow. That case corresponds to $\Theta = 0$. The transit time of sound between two transducers A and B can be found through the average fluid velocity v_c:

$$T = \frac{D}{c \pm v_c \cos \Theta}, \tag{11.31}$$

where c is the velocity of sound in the fluid. The plus/minus signs refer to the downstream/upstream directions, respectively. The velocity v_c is the flow velocity averaged along the path of the ultrasound. Gessner [12] has shown that for laminar

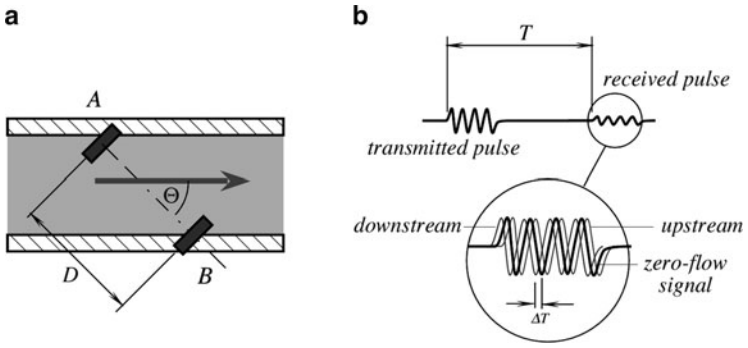

Fig. 11.13 Ultrasonic flowmeter position of transmitter–receiver crystals in the flow (**a**); waveforms in the circuit (**b**)

flow $v_c = 4v_a/3$, and for turbulent flow, $v_c = 1.07v_a$, where v_a is the flow averaged over the cross-sectional area. By taking the difference between the downstream and upstream velocities we find [13]

$$\Delta T = \frac{2Dv_c \cos\Theta}{c^2 + v_c \cos^2\Theta} \approx \frac{2Dv_c \cos\Theta}{c^2}, \tag{11.32}$$

which is true for the most practical cases when $c \gg v_c\cos\Theta$. To improve the signal-to-noise ratio, the transit time is often measured for both upstream and downstream directions. That is, each piezoelectric crystal at one time works as a transmitter and at the other time as a receiver. This can be accomplished by a selector (Fig. 11.14), which is clocked by a relatively slow sampling rate (400 Hz in this example). The sinusoidal ultrasonic waves (about 3 MHz) are transmitted as bursts with the same slow clock rate (400 Hz). A received sinusoidal burst is delayed from the transmitted one by time T, which is modulated by the flow (Fig. 11.13b). This time is detected by a transit time detector, then, the time difference in both directions is recovered by a synchronous detector. Such a system can achieve a quite good accuracy, with a zero-drift as small as 5×10^3 m/s^2 over the 4-h period.

An alternative way to measure flow with the ultrasonic sensors is to detect a phase difference in transmitted and received pulses in the up- and downstream directions. The phase differential can be derived from (11.32)

$$\Delta\varphi = \frac{4fDv_c \cos\Theta}{c^2}, \tag{11.33}$$

where f is the ultrasonic frequency. It is clear that the sensitivity is better with the increase in the frequency, however, at higher frequencies one should expect stronger sound attenuation in the system, which may cause reduction in the signal-to-noise ratio.

For the Doppler flow measurements, continuous ultrasonic waves can be used. Figure 11.15 shows a flowmeter with a transmitter–receiver assembly positioned inside the flowing stream. Like in a Doppler radio receiver, transmitted and received frequencies are mixed in a nonlinear circuit (a mixer). The output low

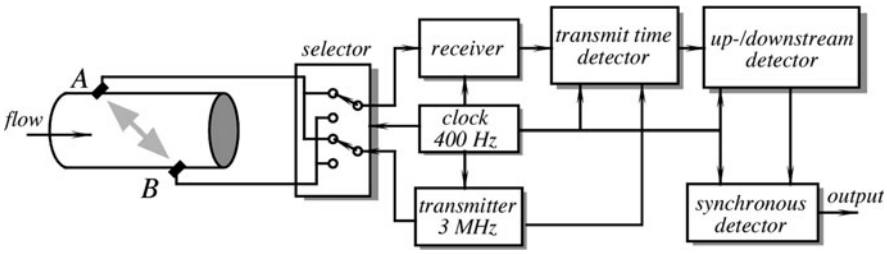

Fig. 11.14 Block diagram of an ultrasonic flowmeter with alternating transmitter and receiver

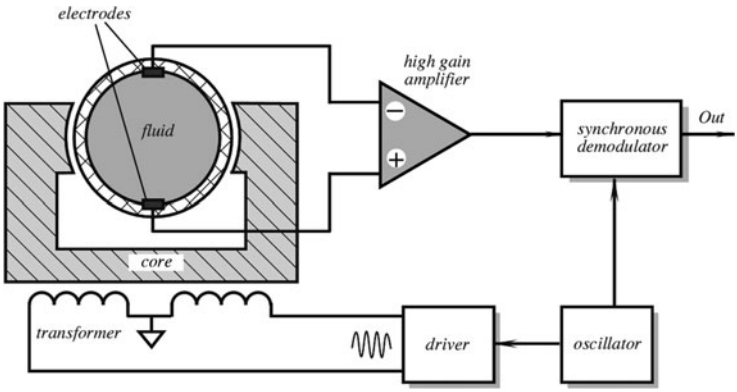

Fig. 11.15 Ultrasonic Doppler flowmeter

frequency differential harmonics are selected by a bandpass filter. That differential is defined as

$$\Delta f = f_s - f_r \approx \pm \frac{2 f_s v}{c}, \tag{11.34}$$

where f_s and f_r are the frequencies in the transmitting and receiving crystals respectively, and the plus/minus signs indicate different directions of flow. An important conclusion from the above equation is that the differential frequency is directly proportional to the flow velocity. Obviously, the piezoelectric crystals must have much smaller sizes than the clearance of the tube of flow. Hence, the measured velocity is not the average but rather a localized velocity of flow. In practical systems, it is desirable to calibrate ultrasonic sensors with actual fluids over the useful temperature range, so that contribution of a fluid viscosity is accounted for.

An ultrasonic piezoelectric sensors/transducers can be fabricated of small ceramic discs encapsulated into a flowmeter body. The surface of the crystal can be protected by a suitable material, for instance, silicone rubber. An obvious advantage of an ultrasonic sensor is in its ability to measure flow without a direct contact with the fluid.

11.5 Electromagnetic Sensors

The electromagnetic flow sensors are useful for measuring the movement of conductive liquids. The operating principle is based on the discovery of Faraday and Henry (see Sect. 3.4) of the electromagnetic induction. When a conductive media, wire, for instance, or for this particular purpose, flowing conductive liquid crosses the magnetic flux lines, e.m.f. is generated in the conductor. As it follows from (3.37), the value of e.m.f. is proportional to velocity of moving conductor.

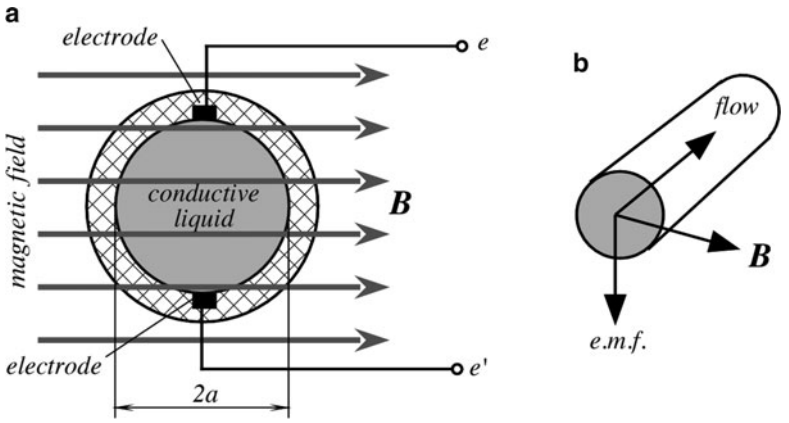

Fig. 11.16 Principle of electromagnetic flowmeter position of electrodes is perpendicular to the magnetic field (**a**); relationships between flow and electrical and magnetic vectors (**b**)

Figure 11.16 illustrates a tube of flow positioned into the magnetic field $\boldsymbol{B}$. There are two electrodes incorporated into a tube to pick up e.m.f. induced in the liquid. The magnitude of the e.m.f. is defined by

$$\mathrm{V} = e - e = 2a\mathbf{B}v, \tag{11.35}$$

where a is the radius of the tube of flow and ν is the velocity of flow.

By solving the Maxwell's equations, it can be shown that for a typical case when the fluid velocity is nonuniform within the cross-sectional area but remains symmetrical about the tube axis (axisymmetrical), the e.m.f generated is the same as that given above, except that ν is replaced by the average velocity, v_{a} (11.3):

$$v_a = \frac{1}{\pi a^2}\int_0^a 2\pi v r \, dr, \tag{11.36}$$

where r is the distance from the center of the tube. Equation (11.35) can be expressed in terms of the volumetric flow rate

$$v = \frac{2\Lambda\mathbf{B}}{\pi a}. \tag{11.37}$$

It follows from the above equation that the voltage registered across the pick-up electrodes is independent of the flow profile or fluid conductivity. For a given tube geometry and the magnetic flux, it depends only on the instantaneous volumetric flow rate.

There are two general methods of inducing voltage in the pick-up electrodes. The first is a dc method where the magnetic flux density is constant and induced voltage is a dc or slow changing signal. One problem associated with this method is a polarization of the electrodes due to small but unidirectional current passing

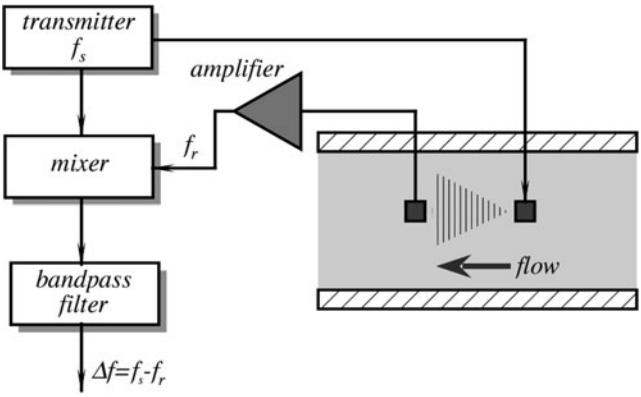

Fig. 11.17 Electromagnetic flowmeter with synchronous (phase sensitive) demodulator

through their surface. The other problem is a low frequency noise, which makes it difficult to detect small flow rates.

Another and far better method of excitation is with an alternating magnetic field, which causes appearance of an ac voltage across the electrodes (Fig. 11.17). Naturally, the frequency of the magnetic field should meet a condition of the Nyquist rate. That is, it must be at least two times higher than the highest frequency of flow rate spectrum variations. In practice, the excitation frequency is selected in the range between 100 and 1,000 Hz.

11.6 Breeze Sensor

In some applications, it is desirable just to merely detect a change in the air (or any other gas for that matter) movement, rather than to measure its flow rate quantitatively. This task can be accomplished by a breeze sensor, which produces an output transient whenever the velocity of the gas flow happens to change. One example of such a device is a piezoelectric breeze sensor produced by Nippon Ceramic, Japan. A sensor contains a pair of the piezoelectric (or pyroelectric) elements,[3] where one is exposed to ambient air and the other is protected by the encapsulating resin coating. Two sensors are required for a differential compensation of variations in ambient temperature. The elements are connected in a series-opposed circuit, that is, whenever both of them generate the same electric charge, the resulting voltage

[3]In this sensor, the crystalline element, which is poled during the manufacturing process, is the same as used in piezo- or pyroelectric sensors. However, the operating principle of the breeze sensor is neither related to mechanical stress nor heat flow. Nevertheless, for the simplicity of the description, we will use the term piezoelectric.

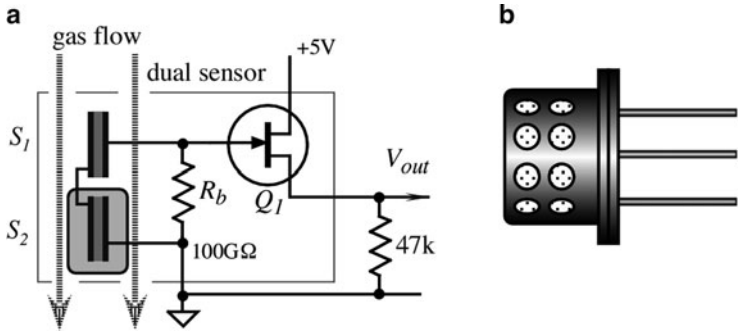

Fig. 11.18 Piezoelectric breeze sensor circuit diagram (**a**); packaging in a TO-5 can (**b**)

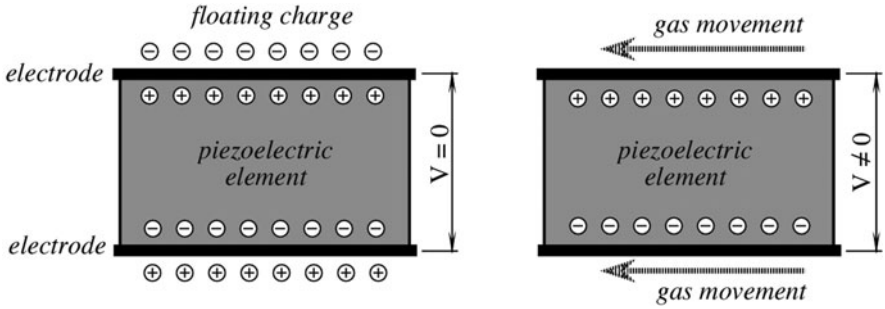

Fig. 11.19 In a breeze sensor, gas movement strips off electric charges from the surface of a piezoelectric element

across the bias resistor R_b (Fig. 11.18a) is essentially zero. Both elements, the bias resistor, and the JFET voltage follower are encapsulated into a TO-5 metal housing with vents for exposing the S_1 element to the gas movement (Fig. 11.18b).

An operating principle of the sensor is illustrated in Fig. 11.19. When airflow is either absent or is very steady, the charge across the piezoelectric element is balanced. The element internal electric dipoles, which were oriented during the poling process (Sect. 3.6.1), are balanced by both the free carriers inside the material and by the charged floating air molecules at the element's surface. In the result, voltage across the piezoelectric elements S_1 and S_2 is zero, which results in the baseline output voltage V_{out}.

When the gas flow across both S_1 surfaces change (S_2 surfaces are protected by resin), the moving gas molecules strip off the floating charges from the element. This results in appearance of a voltage across the element's electrodes because the internally poled dipoles are no longer balanced by the outside floating charges. The voltage is repeated by the JFET follower, which serves as an impedance converter, and appears as a transient in the output terminal.

11.7 Coriolis Mass Flow Sensors

Coriolis flowmeters measure flow of mass directly, as opposed to those that measure velocity or volume [14]. Coriolis flowmeters are virtually unaffected by the fluid pressure, temperature, viscosity, and density. As a result, Coriolis meters can be used without recalibration and without compensating for parameters specific to a particular type of fluid. While these meters were used mainly for liquids when they were first introduced, they have become adaptable for the gas applications.

Coriolis flowmeters are named after Gaspard G. Coriolis (1792–1843), a French civil engineer and physicist. A Coriolis sensor typically consists of one or two vibrating tubes with an inlet and an outlet. A typical material for the tube is stainless steel. It is critical for the meter accuracy to prevent any mechanical or chemical attack of the tube or its lining by the flowing fluid. Some tubes are U-shaped but a wide variety of shapes have been also employed. The thinner tubes are used for gas while thicker tubes are more appropriate for liquids. The Coriolis tube is set to vibration by an auxiliary electromechanical drive system.

Fluid enters the meter in the inlet. A mass flow is determined based on the action of the fluid on the vibrating tubes. As fluid moves from the inlet to outlet, it develops different forces depending on its acceleration that is the result of the tube vibration.

The Coriolis force induced by the flow is described by the following equation:

$$\boldsymbol{F} = 2m\omega\boldsymbol{v}, \tag{11.38}$$

where m is the mass, ω is the rotating circular frequency, and $\boldsymbol{v}$ is the vector of the average fluid velocity. As a result of these forces, the tube takes on a twisting motion as it passes though the vibrating cycle. The amount of twist is directly proportional to the mass flow through the tube. Figure 11.20a shows the Coriolis flow tube in a no-flow situation, and Fig. 11.20b shows Coriolis tube with the flow.

At a no-flow state, the tube vibrates identically at it inlet and outlet sides with the sine-wave motions with the zero phase shift between them. During flow, the tube twists in response to the flow, and the inlet and outlet sides vibrate differently with a phase shift between them (Fig. 11.20c). The main disadvantage of the Coriolis sensor its relatively high initial cost. However, the versatility of Coriolis sensors in

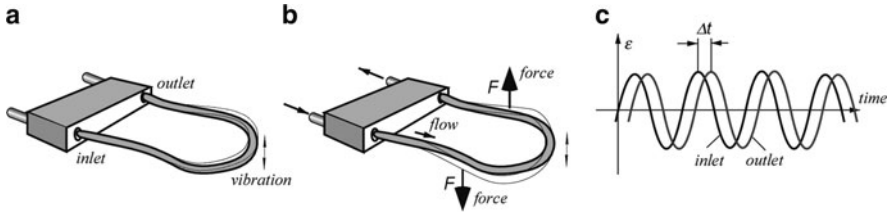

Fig. 11.20 Coriolis tube with no flow (**a**); twist of the tube with flow (**b**); vibrating phase shift resulted from Coriolis forces (**c**)

handling multiple fluids makes them very useful for the plants where the flow of multiple fluid types must be measured. There are also an increasing number of the gas applications for the Coriolis meters.

11.8 Drag Force Sensors

When fluid motion is sporadic, multidirectional, and turbulent, a drag force flow sensor may be quite efficient. Application of such flowmeters include environmental monitoring, meteorology, hydrology, and maritime studies to measure speeds of air or water flow and turbulence close to surface [15]. In the flowmeter, a solid object known as a drag element or target is exposed to the flow of fluid. The force exerted by the fluid on the drag element is measured and converted to an electrical signal indicative of value for speed of flow. An important advantage of the drag sensor is that it can be made to generate a measurement of flow in two dimensions, or even in three dimensions, as well as of flow speed. To implement feature, the drag element must be symmetrical in the appropriate number of dimensions. These flowmeters have been used by industry, utilities, aerospace, and research laboratories to measure the flow of uni- and bi-directional liquids (including cryogenic), gases, and steam (both saturated and superheated) for nearly half a century.

The operation of the sensor is based on strain measurement of deformation of an elastic rubber cantilever to which a force is applied by a spherical symmetrical drag element (Fig. 11.21). An ideal drag element however is a flat disk [16], because this configuration gives a drag coefficient independent of the flow rate. Using a spherical drag element, which departs from the ideal of a flat disk, the drag coefficient may vary with flow rate, and therefore the gauge must be calibrated and optimized for the conditions of intended use. The strain measurement can be performed with

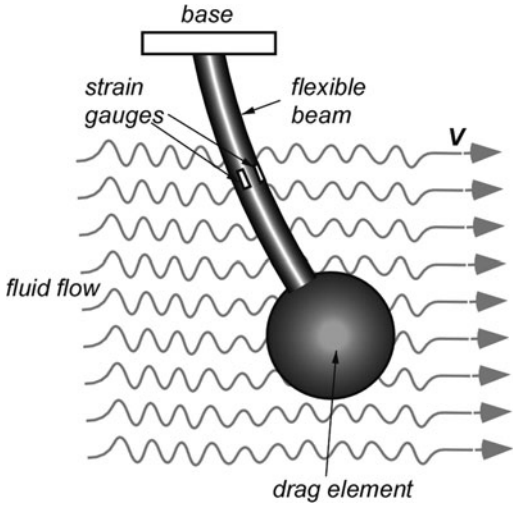

Fig. 11.21 Drag force sensor

strain gages that should be physically protected from interaction with moving fluids.

The drag force F, exerted by incompressible fluid on a solid object exposed to it is given by the drag equation:

$$F_D = C_D \rho A V^2, \tag{11.39}$$

where ρ is fluid density, V is fluid velocity at the point of measurement, A is projected area of the body normal to the flow, and C_D is the overall drag coefficient. C_D is a dimensionless factor, whose magnitude depends primarily on the physical shape of the object and its orientation relative to the fluid stream. If mass of the supporting beam is ignored, the developed strain is

$$\varepsilon = \frac{3 C_D \rho A V^2 (L - x)}{E a^2 b}, \tag{11.40}$$

where L is the beam length, x is the point coordinate on the beam where the strain gauges are located, E is the Young's modulus of elasticity, a and b are the geometry target factors. It is seen that the strain in a beam is a square law function of the fluid flow rate.

11.9 Dust and Smoke Detectors

Smoke and air gas impurity sensors are intended for detecting presence of small airborne particles and have wide range of applications. Even though these detectors do not monitor air flow, their operation essentially requires movement of gas through the detection chamber of the sensor. By far the most popular are the smoke detectors. Such a detector is positioned on or near a ceiling (Fig. 11.22a). It has an inlet for air that can flow through the detector passively or can be drown by a forced convection with help from a fan or blower. Airborne particles greatly vary in size depending on their origin. Table 11.1 exemplifies some contaminants that either may present health hazard or are manifestations of troubling events (e.g., a fire). To detect presence of small particles suspended in air, nowadays two types of sensors are widely employed: ionization and optical detectors.

11.9.1 Ionization Detector

This detector is especially useful for detecting smoke composed of very small particles (submicron) like those generated by large hot fires. The key part of this type of sensor is an ionization chamber containing less than a milligram of

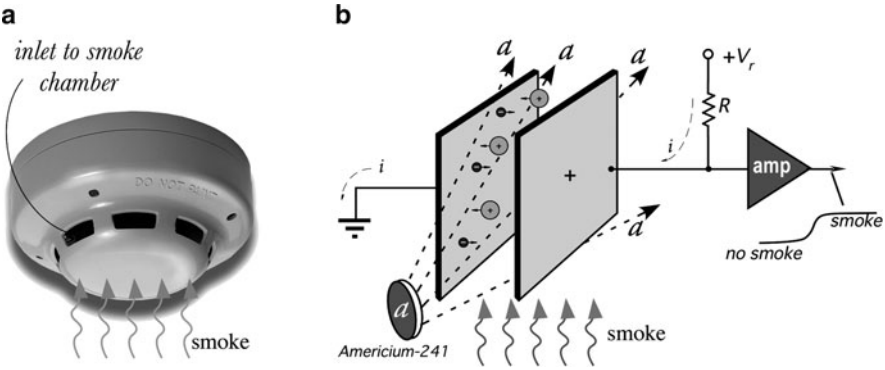

Fig. 11.22 Smoke detector (**a**) and a concept of the ionization smoke sensor (**b**)

Table 11.1 Sizes of some airborne contaminants

Particle	Particle size (μm)	Particle	Particle size (μm)
Glass wool	1,000	Coal dust	1–100
Spanish moss pollen	150–750	Smoke from synthetic materials	1–50
Beach sand	100–10,000	Face powder	0.1–30
Mist	70–350	Asbestos	0.7–90
Pollens	10–1,000	Calcium zinc dust	0.7–20
Textile fibers	10–1,000	Paint pigments	0.1–5
Fiberglass insulation	1–1,000	Car emission	1–150
Grain dusts	5–1,000	Clay	0.1–50
Human hair	40–300	Humidifier	0.9–3
Dust mites	100–300	Copier toner	0.5–15
Saw dust	30–600	Liquid droplets	0.5–5
Cement dust	3–100	Insecticide dusts	0.5–10
Mold spores	10–30	Anthrax	1–5
Textile dust	6–20	Yeast cells	1–50
Spider web	2–3	Carbon black dust	0.2–10
Spores	3–40	Atmospheric dust	0.001–40
Combustion-related carbon monoxide from motor vehicles, wood burning, open burning, industrial processes	Up to 2.5	Smoldering or flaming cooking oil	0.03–0.9
Sea salt	0.035–0.5	Combustion	0.01–0.1
Bacteria	0.3–60	Smoke from natural materials	0.01–0.1
Burning wood	0.2–3	Tobacco smoke	0.01–4
Coal flue gas	0.08–0.2	Viruses	0.005–0.3
Oil smoke	0.03–1	Pesticides and herbicides	0.001

radioactive element Americium-241 (Am^{241}). This element is a natural source of alfa-particles resulted from the so-called alfa decay. The ionization chamber resembles a capacitor with two opposite electrodes (Fig. 11.22b) having shapes either of parallel plates or a coaxial cylinder, where one plate is electrically connected to ground (or the negative side of the power source) and the other, through resistor R, is connected to a positive voltage $+V_r$ (few volts) [17]. The voltage creates an electric field between the plates. Space between the plates is filled with air drawn from inlets at the sides of the plates.

If alfa-particles are absent, no current can pass from the positive plate to the grounded plate, because air normally is not electrically conductive. Alfa-particles emanated by the radioactive element have kinetic energy about 5 MeV, enough to ionize air molecules by breaking them into positively charged ions and negatively charged electrons. The charged ions and electrons are being pulled by the electric field in the opposite directions; electrons to the positive plate and ions to the grounded plate. This results in small constant electric current i flowing from the voltage source V_r, through resistor R, the air-filled space between the plates, and to the "ground." As a result, the input voltage at the amplifier "amp" drops, indicating that no smoke is present in the ionization chamber.

When smoke is drawn into the ionization chamber between the plates, the smoke particles absorb alfa-radiation, thus reducing air ionization and subsequently current i drops. This increases voltage at input of the amplifier, manifesting presence of smoke inside the ionization chamber. Since Americium-241 has a half-life of 432.2 years, the life time of the ionization source is long enough for all practical purposes.

The reasons why alpha radiation is used instead of beta or gamma is twofold: a higher ability of alfa radiation to ionize air and low penetrating power of the alfa particles,[4] so the radiation will be absorbed by the smoke detector housing, reducing a potential harm to humans.

11.9.2 Optical Detector

Another type of a smoke or dust detector is based on measuring scattering of light (see Sect. 3.13.2). An optical detector includes a light emitter (incandescent bulb, infrared LED or a laser diode) and a photosensor, usually a photodiode or phototransistor (Fig. 11.23). The light emitter and detector are positioned inside a light-tight enclosure in such a way as to prevent any photons reaching the detector directly from the emitter or by reflections from the enclosure inner walls. The enclosure should also protect the photosensor from ambient light. To achieve these difficult requirements, the light emitter and photosensor are positioned inside the

[4]Alfa radiation consists of Helium-4 positively charged nucleus and due to high mass travels with the speed of only about 15,000 km/s. Thus, it easily can be stopped by just a thin tissue paper and, due to collisions with air molecules, travels in air at distances no farther than few centimeters.

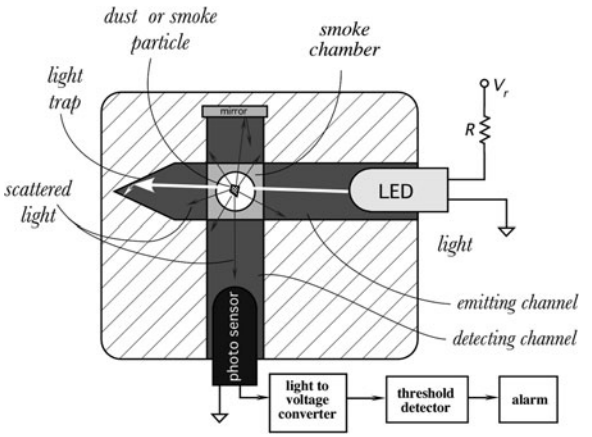

Fig. 11.23 Optical smoke detector

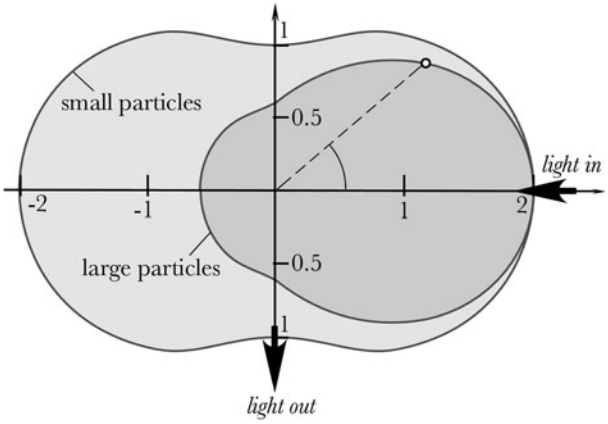

Fig. 11.24 Scattering directional diagram

individual emitting and detecting channels which cross each other preferably at a 90° angle. This is not an optimal angle since light scattering is not the strongest at 90° (Fig. 11.24) but this is the best angle for minimizing a chance for stray light passing from the photo emitter to the photo sensor. Many practical optical smoke detectors [18] use smaller angles on the expense of a larger size and mechanical complexity: to trap the unwanted light before it reaches a photo sensor.

The interior walls of the channels have low light reflectivity surfaces. In addition, the emitting channel has the far end that is conically shaped to prevent a spurious light reflection toward the photo sensor. This shape is called a light trap. The emitter and photo sensor may have built-in lenses to shape the narrow-angle beams, thus assuring that very little light strikes the chamber walls. The space where two channels cross is called a "scattering chamber" or "smoke chamber,"

which is open to ambient air or connected to the source of tested air. The smoke chamber has the inlets and outlets for air passing through, but shielded from ambient light.

In presence of clean air inside the scattering chamber, light beams from the emitter cannot reach the photo sensor (in empty space, light cannot go around the corner) and thus the photo sensor produces a very low output current, called the dark current (see Sect. 14.2). When dust or smoke enters the scattering chamber, it appears across the path of the light beam and some light is scattered by the particles in all directions (Fig. 11.24), including a direction toward the photo sensor. Smaller airborne particles cause the Rayleigh scattering while the larger particles cause specular reflections in many directions (Fig. 3.49), including the direction along the detecting channel toward the photo sensor. The detecting channel may have a mirror at the opposite end to bounce some scattered light toward the photo sensor, thus increasing the sensor's sensitivity. Whatever is the physical nature of scattering, thanks to air impurities the photo sensor now sees some light "in the tunnel." Small particles being much more numerous create a relatively constant shift in the photo current, while the larger particles will glitter and result in a pulsing photo current. The effect of appearance of light in the detecting channel resembles appearance of light rays when the sun is partially covered by clouds: the sunlight rays passing through openings in the cloud become visible due to light scattering by water droplets and airborne dust particles.

A photodiode along with the interface circuit serves as light-to-voltage converter (see Sect. 5.3) whose output is fed to a threshold detector connected to an alarm. To respond to the light pulses resulted from larger particles, the electronic interface circuit should have a sufficiently wide frequency bandwidth. The optical smoke detectors are less sensitive to false alarms resulting from steam or cooking fumes in kitchen or steam from the bathroom than the ionization smoke alarms. They are especially efficient for detecting smoke from smoldering fires.

References

1. Benedict RP (1984) Fundamentals of temperature, pressure, and flow measurements, 3rd edn. Wiley, New York
2. Cho ST, Wise KD (1991) A high performance microflowmeter with built-in self test. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers, pp 400–403, IEEE, 1991
3. Bruun HH (1995) Hot-wire anemometry. Principles and signal analysis. Oxford Science, Oxford
4. King LV (1914) On the convention of heat from small cylinders in a stream of fluid. Philos Trans R Soc A214:373
5. Collis DC, Williams MJ (1959) Two-dimensional convection from heated wires at low Reynolds' numbers. J Fluid Mech 6:357
6. Fraden J, Rutstein A (2007) Clogging detector for air filter. US Patent 7178410, 20 Feb
7. Fraden J (2009) Detector of low levels of gas pressure and flow. US Patent 7490512, 17 Feb

8. Van Herwaarden AW, Sarro PM (1986) Thermal sensors based on the Seebeck effect. Sens Actuators 10:321–346
9. Chiu N-F et al. (2005) Low power consumption design of micro-machined thermal sensor for portable spirometer. Tamkang J Sci Eng 8(3):225–230
10. Wachutka G, Lenggenhager R, Moser D, Baltes H (1991) Analytical 2D-model of CMOS micromachined gas flow sensors. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers. ©IEEE
11. Esashi M (1991) Micro flow sensor and integrated magnetic oxygen sensor using it. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers. IEEE
12. Gessner U (1969) The performance of the ultrasonic flowmeter in complex velocity profiles. IEEE Trans Bio-Med Eng MBE-16 16:139–142
13. Cobbold RSC (1974) Transducers for biomedical measurements. Wiley, New York
14. Yoder J (2000) Coriolis effect mass flowmeters. In: Webster J (ed) Mechanical variables measurement. CRC Press LLC, Boca Raton, FL
15. Philip-Chandy R, Morgan R, Scully PJ (2000) Drag force flowmeters. In: Webster J (ed) Mechanical variables measurement. CRC Press LLC, Boca Raton, FL
16. Clarke T (1986) Design and operation of target flowmeters. In: Encyclopedia of fluid mechanics, vol 1. Gulf Publishing Company, Houston, TX
17. Dobrzanski J, Gardner EB (1977) Ionization smoke detector and alarm system. US Patent 4037206, 19 July
18. Steele DF, Enmark RB (1975) Optical smoke detector. US Patent 3863076, 28 Jan

Chapter 12
Acoustic Sensors

Sound is the vocabulary of nature

– Pierre Schaeffer, French composer

The fundamentals of acoustics are given in Sect. 3.10. Here we will discuss the acoustic sensors for various frequency ranges. The audible range sensors are generally called the microphones, however, the name is often used even for the ultrasonic and infrasonic waves. In essence, a microphone is a pressure transducer adapted for transduction of sound waves over a broad spectral range, which generally excludes very low frequencies below few Hz. The microphones differ by their sensitivity, directional characteristics, frequency bandwidth, dynamic range, sizes, etc. Also, their designs are quite different depending on the media from which sound waves are sensed. For example, for perception of air waves or vibrations in solids, the sensor is called a microphone, while for operation in liquids, it is called a hydrophone (even if liquid is not water – from the Greek name of the mythological water serpent *Hydra*). The main difference between a pressure sensor and an acoustic sensor is that latter does not need to measure constant or very slow changing pressures. Its operating frequency range usually starts at several hertz (or as low as tens of millihertz for some applications), while the upper operating frequency limit is quite high – up to several megahertz for the ultrasonic applications and even gigahertz in a surface acoustic wave (SAW) device.

Since acoustic waves are mechanical pressure waves, any microphone or hydrophone has the same basic structure as a pressure sensor: it is comprised of a moving diaphragm and a displacement transducer, which converts the diaphragm's deflections into an electrical signal. All microphones and hydrophones differ by the designs of these two essential components. Also, they may include some additional parts such as mufflers, focusing reflectors or lenses, etc., however, in this chapter we will review only the sensing parts of some of the most interesting, from our point of view, acoustic sensors.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_12, © Springer Science+Business Media, LLC 2010

12.1 Resistive Microphones

In the past, resistive pressure converters (pressure to electricity) were used quite extensively in microphones. The converter consisted of a semiconductive powder (usually graphite) whose bulk resistivity was sensitive to pressure. Nowadays we would say that the powder possessed piezoresistive properties. However, these early devices had quite a limited dynamic range, poor frequency response, and a high noise floor. A carbon microphone is a capsule containing carbon granules pressed between two metal plates. A voltage is applied across the metal plates, causing a small current to flow through the carbon. One of the plates, the diaphragm, vibrates under the incident sound waves, applying a varying pressure to the carbon. The changing pressure deforms the granules, causing the contact area between each pair of adjacent granules to change, and this causes the electrical resistance of the mass of granules to change. The changes in resistance cause a corresponding change in voltage across the two plates, and hence in the current flowing through the microphone, producing the electrical signal. Carbon microphones were once commonly used in telephones.

Presently, the same piezoresistive principle can be employed in the micromachined sensors, where stress sensitive resistors are the integral parts of a silicon diaphragm (Sect. 10.5).

12.2 Condenser Microphones

If a parallel-plate capacitor is given an electric charge, q, voltage across its plates is governed by (3.19). On the other hand, according to (3.20) the capacitance depends on distance d between the plates. Thus solving these two equations for voltage we arrive at

$$V = q\frac{d}{A\varepsilon_0}, \tag{12.1}$$

where $\varepsilon_0 = 8.8542 \times 10^{-12}$ C^2/Nm^2 is the permittivity constant (Sect. 3.1). The above equation is the basis for operation of the condenser microphones, which is the other way to say "capacitive" microphones. Thus, a capacitive microphone linearly converts a distance between the plates into electrical voltage, which can be further amplified. The device essentially requires a source of an electric charge q whose magnitude directly determines the microphone sensitivity. The charge can be provided either from an external power supply having a voltage in the range from 20 to 200 V, or from an internal source capable of producing such a charge. This is accomplished by a built-in electret layer, which is a polarized dielectric crystal.

Presently, many condenser microphones are fabricated with silicon diaphragms, which serve two purposes: to convert acoustic pressure into displacement, and to

act as a moving plate of a capacitor. Some promising designs are described in [1–3]. To achieve high sensitivity, a bias voltage should be as large as possible, resulting in a large static deflection of the diaphragm, which may result in a reduced shock resistivity and lower dynamic range. Besides, if the air gap between the diaphragm and the backplate is very small, the acoustic resistance of the air gap will reduce the mechanical sensitivity of the microphone at higher frequencies. For instance, at an air gap of 2 μm, an upper cutoff frequency of only 2 kHz has been measured [1].

One way to improve the characteristics of a condenser microphone is to use a mechanical feedback from the output of the amplifier to the diaphragm [4]. Figure 12.1a shows a circuit diagram and Fig. 12.1b is a drawing of interdigitized electrodes of the microphone. The electrodes serve different purposes: one is for the conversion of a diaphragm displacement into voltage at the input of the amplifier A_1 while the other electrode is for converting feedback voltage V_a into a mechanical deflection by means of electrostatic force. The mechanical feedback clearly improves linearity and the frequency range of the microphone, however, it significantly reduces the deflection which results in a lower sensitivity.

Radiofrequency (RF) condenser microphones use additional RF signal generated by a low-noise oscillator. The oscillator may either be frequency modulated by the capacitance changes produced by the sound waves moving the capsule diaphragm, or the capsule may be part of a resonant circuit that modulates the amplitude of the fixed-frequency oscillator signal. Demodulation yields a low-noise audio frequency signal with a very low source impedance. This technique permits the use of a diaphragm with looser tension, which may be used to achieve wider frequency response due to higher compliance.

Condenser microphones span the range from telephone transmitters to inexpensive karaoke microphones to high-fidelity recording microphones. They generally produce a high-quality audio signal and are now the popular choice in laboratory and studio recording applications. The inherent suitability of this technology is due to the very small mass that must be moved by the incident sound wave. For further reading on condenser microphones, an excellent book may be recommended [5].

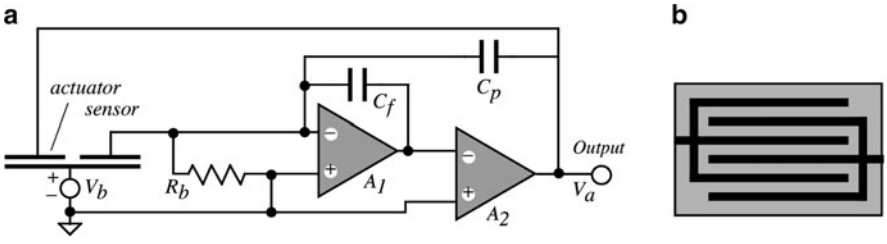

Fig. 12.1 Condenser microphone with a mechanical feedback Circuit diagram (**a**); interdigitized electrodes on the diaphragm (**b**) (adapted from [4])

12.3 Fiber-Optic Microphone

Direct acoustic measurements in hostile environments, such as in turbojets or rocket engines, require sensors, which can withstand high heat and strong vibrations. The acoustic measurements under such hard conditions are required for computational fluid dynamics (CFD) code validation, structural acoustic tests, and jet noise abatement. For such applications, a fiber-optic interferometric microphone can be quite suitable. One such design [6] is comprised of a single-mode temperature-insensitive Michelson interferometer and a reflective plate diaphragm. The interferometer monitors the plate deflection, which is directly related to the acoustic pressure. The sensor is water cooled to provide thermal protection for the optical materials and to stabilize the mechanical properties of the diaphragm.

To provide an effect of interference between the incoming and outgoing light beams: two fibers are fused together and cleaved at the minimum tapered region (Fig. 12.2). The fibers are incorporated into a stainless steel tube, which is water cooled. The internal space in the tube is filled with epoxy, while the end of the tube is polished until the optical fibers are observed. Next, aluminum is selectively deposited to one of the fused fiber core ends to make its surface mirror reflective. This fiber serves as a reference arm of the microphone. The other fiber core is left open and serves as the sensing arm. Temperature insensitivity is obtained by the close proximity of the reference and sensing arms of the assembly.

Light from a laser source (a laser diode operating near 1.3 μm wavelength) enters one of the cores and propagates toward the fused end, where it is coupled to the other fiber core. When reaching the end of the core, light in the reference core is reflected from the aluminum mirror toward the input and output sides of the sensor. The portion of light, which goes toward the input is lost and makes no effect on the measurement, while the portion which goes to the output, strikes the detector's surface. That portion of light that travels to the right in the sensing core exits the fiber and strikes the copper diaphragm. Part of the light is reflected from the diaphragm back toward the sensing fiber and propagates to the output end,

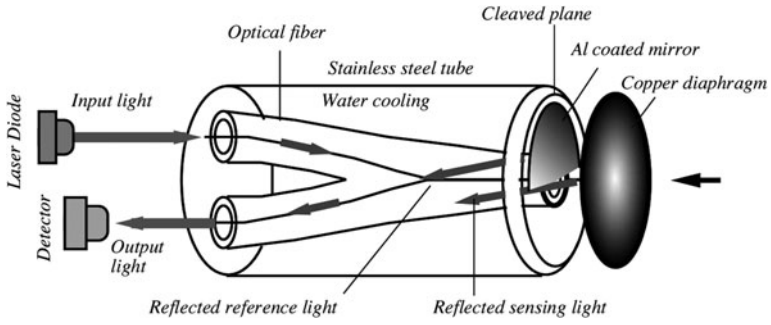

Fig. 12.2 Fiber optic interferometric microphone movement of copper diaphragm is converted into light intensity in the detector

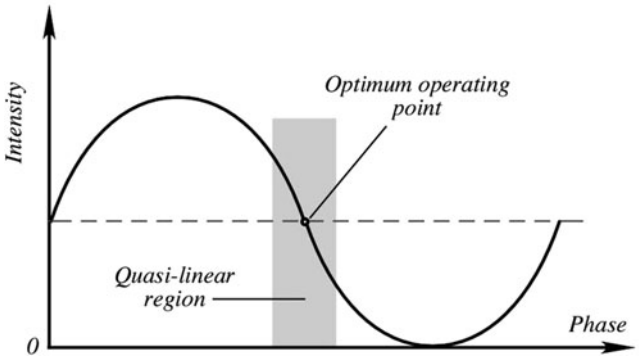

Fig. 12.3 Intensity plot as function of a reflected light phase

along with the reference light. Depending on the position of the diaphragm, the phase of the reflected light will vary, thus becoming different from the phase of the reference light.

While traveling together to the output detector, the reference and sensing lights interfere with one another, resulting in the light intensity modulation. Therefore, the microphone converts the diaphragm displacement into a light intensity. Theoretically, the signal-to-noise ratio in such a sensor is obtainable on the order of 70–80 dB, thus resulting in an average minimum detectable diaphragm displacement of 1 Å (10^{-10} m).

Figure 12.3 shows a typical plot of the optical intensity in the detector versus the phase for the interference patterns. To assure a linear transfer function, the operating point should be selected near the middle of the intensity, where the slope is the highest and the linearity is the best. The slope and the operating point may be changed by adjusting the wavelength of the laser diode. It is important for the deflection to stay within a quarter of the operating wavelength to maintain a proportional input.

The diaphragm is fabricated from a 0.05 mm foil with a 1.25 mm diameter. Copper is selected for the diaphragm due to its good thermal conductivity and relatively low modulus of elasticity. The latter feature allows the use a thicker diaphragm, which provides better heat removal while maintaining a usable natural frequency and deflection. A pressure of 1.4 kPa produces a maximum center deflection of 39 nm (390 Å), which is well within a 1/4 of the operating wavelength (1,300 nm). The maximum acoustic frequency, which can be transferred with the optical microphone, is limited to about 100 kHz, which is well above the desired working range needed for the structural acoustic testing.

12.4 Piezoelectric Microphones

The piezoelectric effect can be used for the design of simple microphones. A piezoelectric crystal is a direct converter of a mechanical stress into an electric

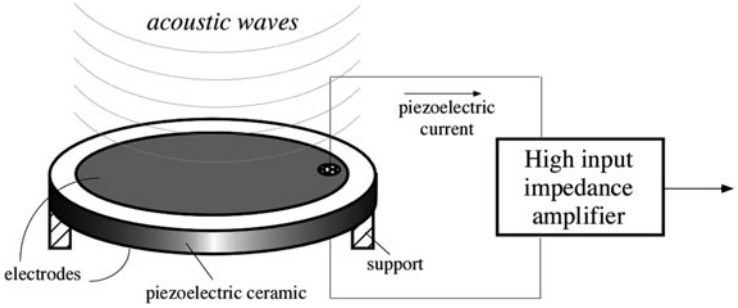

Fig. 12.4 Piezoelectric microphone

charge. The most frequently used material for the sensor is a piezoelectric ceramic, which can operate up to a very high-frequency limit. This is the reason why piezoelectric sensors are used for transduction of ultrasonic waves (Sect. 7.5). Still, even for the audible range, the piezoelectric microphones are used quite extensively. Typical applications are the voice activated devices and blood-pressure measurement apparatuses where the arterial Korotkoff sounds have to be detected. For such acoustically nondemanding applications, the piezoelectric microphone design is quite simple (Fig. 12.4). It consists of a piezoelectric ceramic disk with two electrodes deposited on the opposite sides. The electrodes are connected to wires either by electrically conductive epoxy or by soldering. Since the output electrical impedance of such a microphone is very large, a high input impedance amplifier is required.

Piezoelectric films (PVDF and copolymers) were used for many years as very efficient acoustic pickups in musical instruments [7]. One of the first applications for piezoelectric film was as an acoustic pickup for a violin. Later, the film was introduced for a line of acoustic guitars as a saddle-mounted bridge pickup, mounted in the guitar bridge. The very high fidelity of the pickup led the way to a family of vibration sensing and accelerometer applications. In one guitar pickup, a thick film, compressive (under the saddle) design; another is a low cost accelerometer, while another is an after market pickup design that is taped to the instrument. Because of the low Q-factor[1] of the material, these transducers do not have the self-resonance of hard ceramic pickups. Shielding can be achieved by a foldover design as shown in Fig. 12.5a. The sensing side is the slightly narrower electrode on the inside of the fold. The foldover technique provides a more sensitive pickup than alternative shielding methods because the shield is formed by one of the electrodes. For application in water, the film can be rolled in tubes and many of such tubes can be connected in parallel (Fig. 12.5b).

[1] Q-factor (quality factor) describes how the resonant bandwidth Δf relates to the center frequency f_r: $Q = f_r/\Delta f$. Q is an indicator of energy losses near the resonant frequency.

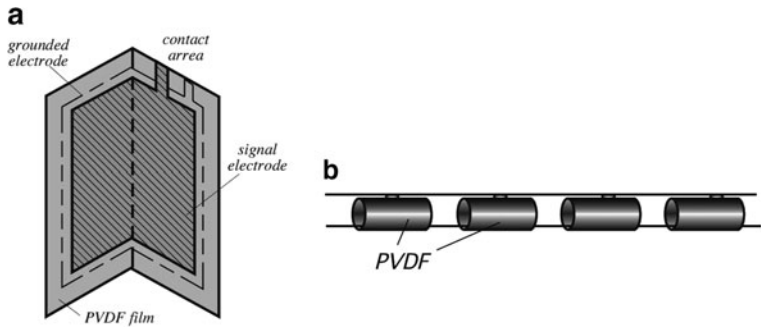

Fig. 12.5 Foldover piezoelectric acoustic pickup (**a**) and arrangement of a piezoelectric film hydrophone (**b**)

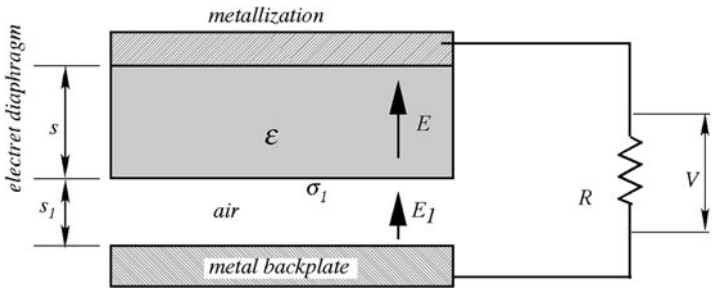

Fig. 12.6 General structure of an electret microphone Thicknesses of layers are exaggerated for clarity (after [9])

12.5 Electret Microphones

An electret is a close relative to piezoelectric and pyroelectric materials. In effect, they are all electrets with enhanced either piezoelectric or pyroelectric properties. An electret is a permanently electrically polarized crystalline dielectric material. The first application of electrets to microphones and earphones where described in 1928 [8]. An electret microphone is an electrostatic transducer consisting of a metallized electret diaphragm and backplate separated from the diaphragm by an air gap (Fig. 12.6).

The upper metallization and a metal backplate are connected through a resistor R voltage V across which can be amplified and used as an output signal. Since the electret is permanently electrically polarized dielectric, charge density σ_1 on its surface is constant and sets in the air gap an electric field E_1. When acoustic wave impinges on the diaphragm, the latter deflects downward reducing the air gap

thickness s_1 for a value of Δs. Under open-circuit conditions, the amplitude of a variable portion of the output voltage becomes

$$V = \frac{s\Delta s}{\varepsilon_0(s + \varepsilon s_1)}. \tag{12.2}$$

Thus, the deflected diaphragm generates voltage across the electrodes. That voltage is in phase with the diaphragm deflection. If the sensor has a capacitance C the above equation should be written as

$$V = \frac{s\Delta s}{\varepsilon_0(s + \varepsilon s_1)} \frac{2\pi fRC}{\sqrt{1 + (2\pi fRC)^2}}, \tag{12.3}$$

where f is frequency of sonic waves.

If the restoring forces are due to elasticity of the air cavities behind the diaphragm (effective thickness is s_0) and the tension T of the membrane, its displacement Δs to a sound pressure Δp assuming negligible losses is given by [10]

$$\Delta s = \frac{\Delta p}{(\gamma p_0/s_0) + (8\pi T/A)}, \tag{12.4}$$

where γ is the specific heat ratio, p_0 is the atmospheric pressure, and A is the membrane area. If we define the electret microphone sensitivity as $\delta_m = \Delta V/\Delta p$, then below resonant, it can be expressed as [9]

$$\delta_m = \frac{ss_0\sigma_1}{\varepsilon_0(s + \varepsilon s_1)\gamma p_0}. \tag{12.5}$$

It is seen that the sensitivity does not depend on area. If mass of the membrane is M, then the resonant frequency is defined by

$$f_r = \frac{1}{2\pi}\sqrt{\frac{p_0}{s_0 M}}. \tag{12.6}$$

This frequency should be selected well above the upper frequency of the microphone's operating range.

The electret microphone differs from other similar detectors in the sense that it does not require a dc bias voltage. For a comparable design dimensions and sensitivity, a condenser microphone would require well over 100 V bias. The mechanical tension of the membrane is generally kept at a relatively low value (about 10 Nm^{-1}), so that the restoring force is determined by the air gap compressibility. A membrane may be fabricated of Teflon FEP, which is permanently charged by an electron beam to give it electret properties. The temperature

coefficient of sensitivity of the electret microphones are in the range of 0.03 dB/°C in the temperature range from −10 to +50°C [11].

Foil-electret (diaphragm) microphones have more desirable features than any other microphone types. Among them is very wide frequency range from 10^{-3} Hz and up to hundreds of MHz. They also feature a flat frequency response (within ±1 dB), low harmonic distortion, low vibration sensitivity, good impulse response, and insensitivity to magnetic fields. Sensitivities of electret microphones are in the range of few mV/μbar.

For operation in the infrasonic range, an electret microphone requires a miniature pressure equalization hole on the backplate. When used in the ultrasonic range, the electret is often given an additional bias (like a condenser microphone) in addition to its own polarization.

Electret microphones are high impedance sensors and thus require high input impedance interface electronics. A JFET transistor has been the input of choice for many years. However, recently monolithic amplifiers gain popularity. An example is the LMV1014 (National Semiconductors), which is an audio amplifier with low current consumption (38 μA) that may operate from a small battery power supply ranging from 1.7 to 5 V.

12.6 Dynamic Microphones

Dynamic microphones work via electromagnetic induction. They are robust, relatively inexpensive, and resistant to moisture. This, coupled with their potentially high gain (before feedback is applied), makes them ideal for on-stage use.

Moving-coil microphones use the same dynamic principle as in a loudspeaker, only reversed (Fig. 12.7a). A small movable induction coil, positioned in the

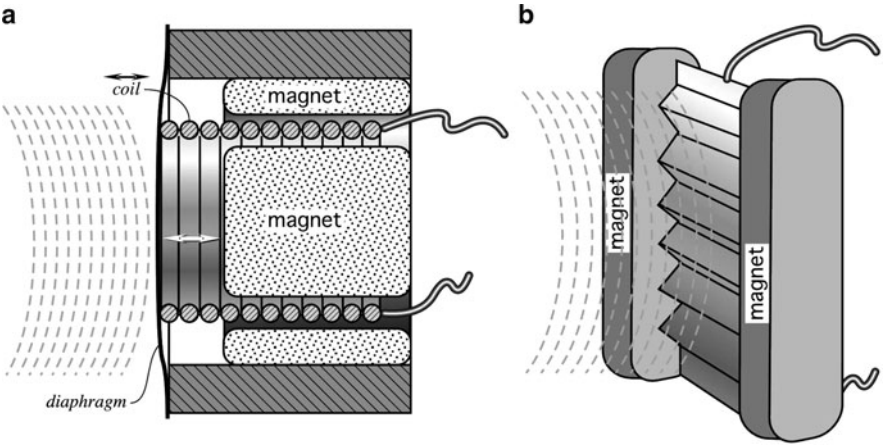

Fig. 12.7 Dynamic microphones: moving coil (**a**) and ribbon (**b**)

magnetic field of a permanent magnet, is attached to the diaphragm. When sound enters through the windscreen of the microphone (not shown in the figure), the sound wave moves the diaphragm. When the diaphragm vibrates, the coil moves in the magnetic field, producing a varying voltage across the coil terminals. This is a result of electromagnetic induction. As it follows from (3.36), a variable magnetic field induces voltage in a coil. Thus, movement of the coil inside a permanent magnet generates the induced voltage and a subsequent current in direct relationship with the rate of changing the magnetic field.

A single dynamic membrane will not respond linearly to all audio frequencies. Some microphones for this reason utilize multiple membranes for the different parts of the audio spectrum and then combine the resulting signals. Combining the multiple signals correctly is difficult and designs that do this are rare and tend to be expensive.

Ribbon microphones use a thin, usually corrugated metal ribbon suspended in a magnetic field (Fig. 12.7b). The ribbon is electrically connected to the microphone's output, and its vibration within the magnetic field generates the electrical signal. Ribbon microphones are similar to moving coil microphones in the sense that both produce sound by means of magnetic induction. Basic ribbon microphones detect sound in a bidirectional (also called figure-eight) pattern because the ribbon, which is open to sound both front and back, responds to the pressure gradient rather than the sound pressure. Though the symmetrical front and rear pickup can be a nuisance in normal stereo recording, the high side rejection can be used to advantage is some applications, especially where a background noise rejection is required.

12.7 Solid-State Acoustic Detectors

Nowadays, use of the acoustic sensors is broader than detecting sound waves in air. Particularly they become increasingly popular for detecting mechanical vibrations in solid for fabrication such sensors as microbalances and SAW devices. Applications range over measuring displacement, concentration of compounds, stress, force, temperature, etc. All such sensors are based on elastic motions in solid parts of the sensor and their major use is serving as parts in other, more complex sensors, for instance, in chemical detectors, accelerometers, pressure sensors, etc. In chemical and biological sensors, the acoustic path, where mechanical waves propagate, may be coated with chemically selective compound, which interact only with the stimulus of interest.

An excitation device (usually of a piezoelectric nature) forces atoms of the solid into vibratory motions about their equilibrium position. The neighboring atoms then produce a restoring force tending to bring the displaced atoms back to their original positions. In the acoustic sensors, vibratory characteristics, such as phase velocity and/or attenuation coefficient, are affected by the stimulus. Thus, in acoustic sensors external stimuli, such as mechanical strain in the sensor's solid, increase the

propagating speed of sound. In other sensors, which are called gravimetric, sorption of molecules or attachment of bacteria cause a reduction of acoustic wave velocity. And in another detectors, called the acoustic viscosity sensors, viscous liquid contacts the active region of an elastic wave sensor and the wave is attenuated.

Acoustic waves propagating in solids have been used quite extensively in electronic devices such as electric filters, delay lines, microactuators, etc. The major advantage of the acoustic waves as compared with electromagnetic waves is their low velocity. Typical velocities in solids range from 1.5×10^3 to 12×10^3 m/s, while the practical SAW utilize the range between 3.8×10^3 to 4.2×10^3 m/s [12]. That is, acoustic velocities are five orders of magnitude smaller than those of electromagnetic waves. This allows for fabrication of miniature sensors operating with frequencies up to 5 GHz.

When the solid-state acoustic sensor is fabricated it is essential to couple the electronic circuit to its mechanical structure where the waves propagate. The most convenient effect to employ is the piezoelectric effect. The effect is reversible (see Sect. 3.6), which means that it works in both ways: the mechanical stress induces electrical polarization charge and the applied electric field stresses the piezoelectric crystal. Thus, the sensor generally has two piezoelectric transducers at both ends: one at the transmitting end for generation of acoustic waves and the other at the receiving end – for conversion of acoustic waves into electrical signal.

Since silicon does not posseses piezoelectric effect, additional piezoelectric material must be deposited on the silicon waver in a form of a thin film [12]. Typical piezoelectric materials used for this purpose are zinc oxide (ZnO), aluminum nitride (AlN), and the so-called solid solution system of lead–zirconite–titanium oxides Pb(Zr, Ti)O_3 known as PZT ceramic. When depositing thin films on the semiconductor material, several major properties must be taken into account. They are as follows:

1. Quality of the adhesion to the substrate
2. Resistance to the external factors (such as fluids which interact with the sensing surface during its operations)
3. Environmental stability (humidity, temperature, mechanical shock and vibration)
4. Value of electromechanical coupling with the substrate
5. Ease of processing by the available technologies
6. Cost

The strength of the piezoelectric effect in elastic-wave devices depends on configuration of the transducting electrodes. Depending on the sensor design, for the bulk excitation (when the waves must propagate through the cross-sectional thickness of the sensor) the electrodes are positioned at the opposite sides and their area is quite large. For the SAW, the excitation electrodes are interdigitized.

Several configurations for the solid-state acoustic sensors are known. They differ by the mode the waves propagate through the material. Figure 12.8 shows two most common versions: a sensor with flextural plate mode (a) and with the acoustic plate mode (b). In the former case, a very thin membrane is flexed by the left pair of the interdigitized electrodes and its vertical deflection induces response in the right pair of the electrodes. As a rule, the membrane thickness is substantially less than the

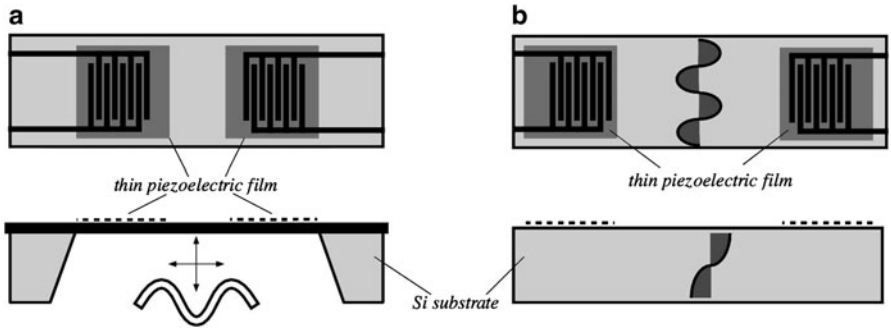

Fig. 12.8 Flextural-plate mode (**a**) and surface acoustic plate mode (**b**) sensors

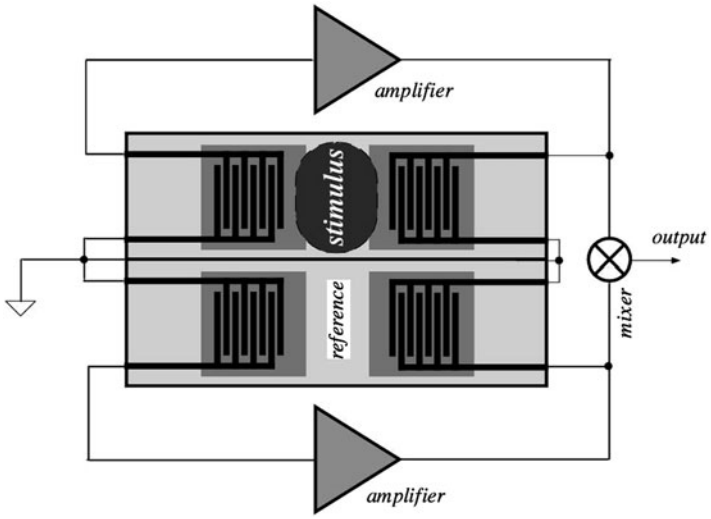

Fig. 12.9 Differential SAW sensor

wavelength of the oscillation. In the latter case, the waves are formed on the surface of a relatively thick plate. In either case, the space between the left and right pairs of the electrodes is used for interaction with the external stimulus, such as pressure, viscous fluid, gaseous molecules, or microscopic particles.

A typical application circuit for SAW includes a SAW plate as a time keeping device of a frequency oscillator. Because many internal and external factors may contribute to propagation of acoustic wave and, subsequently, to change in frequency of oscillation, the determination of change in stimulus may be ambiguous and contain errors. An obvious solution is to use a differential technique, where two identical SAW devices are employed: one device is for sensing the stimulus and the other is reference (Fig. 12.9). The reference device is shielded from stimulus, but subjected to common factors, such as temperature, aging, etc. The difference of the

frequency changes of both oscillators is sensitive only to variations in the stimulus, thus canceling effects of spurious factors.

References

1. Hohm D, Hess G (1989) A subminiature condenser microphone with silicon nitrite membrane and silicon back plate. J Acoust Soc Am 85:476–480
2. Bergqvist J, Rudolf F (1990) A new condenser microphone in silicon. Sens Actuators A21–A23:123–125
3. Sprenkels AJ, Groothengel RA, Verloop AJ, Bergveld P (1989) Development of an electret microphone in silicon. Sens Actuators 17(3&4):509–551
4. van der Donk AGH, Sprenkels AJ, Olthuis W, Bergveld P (1991) Preliminary results of a silicon condenser microphone with internal feedback. In: Transducers'91. International conference on solid-state sensors and actuators. Digest of Technical Papers, pp 262–265, IEEE, 1991
5. Wong SK, Embleton TFW (eds) (1995) AIP handbook of condenser microphones. AIP Press, New York
6. Hellbaum RF et al. (1991) An experimental fiber optic microphone for measurement of acoustic pressure levels in hostile environments. In: Sensors Expo proceedings, 1991, Helmers Publishing, Inc.
7. Piezo film sensors technical manual. Measurement Specialties, Inc., Norristown, PA, http://www.msiusa.com 1999
8. Nishikawa S, Nukijama S (1928) Proc Imp Acad Tokyo 4:290
9. Sessler GM (ed) (1980) Electrets. Springer, Berlin
10. Morse PM (1948) Vibration and sound. McGraw-Hill, New York
11. Griese HJ (1977) Paper Q29. In: Proc 9th Int Conf Acoust, Madrid, 1977
12. Motamedi ME, White RM (1994) Acoustic sensors. In: SMSze (ed) Semiconductor sensors. Wiley, New York, pp 97–151

Chapter 13
Humidity and Moisture Sensors

Water, taken in moderation, cannot hurt anybody

–Mark Twain

13.1 Concept of Humidity

The water content in surrounding air is an important factor for the well-being of humans and animals. The level of comfort is determined by a combination of two factors: relative humidity and ambient temperature. You may be quite comfortable at −30°C (−22°F) in Siberia, where the air is usually very dry in winter, and feel quite miserable in Cleveland near lake Erie at 0°C (+32°F), where air may contain substantial amount of moisture.[1] Humidity is an important factor for operating certain equipments, for instance, high impedance electronic circuits, electrostatic sensitive components, high voltage devices, fine mechanisms, etc. A rule of thumb is to assure a relative humidity near 50% at normal room temperature (20–25°C). This may vary from as low as 38% for the Class-10 clean rooms to 60% in hospital-operating rooms. Moisture is the ingredient common to most manufactured goods and processed materials. It can be said that a significant portion of the U.S. GNP (Gross National Product) is moisture [1].

Humidity can be measured by the instruments called hygrometers. The first hygrometer was invented by Sir John Leslie (1766–1832) [2]. To detect moisture contents, a sensor in a hygrometer must be selective to water, and its internal properties should be modulated by the water concentration. Generally, sensors for moisture, humidity, and dew temperature can be capacitive, conductive, oscillating, or optical. The optical sensors for gases detect dewpoint temperature, while the optical hygrometers for organic solvents employ absorptivity of near-infrared (NIR) light in the spectral range from 1.9 to 2.7 μm [3] (see Fig. 14.15).

[1]Naturally, here we disregard other comfort factors, such as economical, cultural, and political; otherwise, I would not call Siberia a comfortable place.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_13, © Springer Science+Business Media, LLC 2010

There are many ways to express moisture and humidity, often depending on industry or the particular application. Moisture of gases sometimes is expressed in pounds of water vapor per million cubic feet of gas. Moisture in liquids and solids is generally given as a percentage of water per total mass (wet weight basis) but may be given on a dry weight basis. Moisture in liquids with low water miscibility is usually expressed as parts per million by weight (PPM_w).

The term *moisture* generally refers to the water content of any material, but for practical reasons, it is applied only to liquids and solids, while the term humidity is reserved for the water vapor content in gases. Here are some useful definitions.

Moisture – the amount of water contained in a liquid or solid by absorption or adsorption, which can be removed without altering its chemical properties.

Mixing ratio (humidity ratio) r – the mass of water vapor per unit mass of dry gas.

Absolute humidity (mass concentration or density of water vapor) – the mass m of water vapor per unit volume v of wet gas: $d_w = m/v$. In other words, absolute humidity is the density of water vapor component. It can be measured, for example, by passing a measured quantity of air through a moisture-absorbing substance (such as silica-gel) that is weighed before and after the absorption. Absolute humidity is expressed in grams per cubic meter, or in grains per cubic foot. Since this measure is also a function of atmospheric pressure, it is not generally useful in engineering practice.

Relative humidity (*RH*) is the ratio of the actual vapor pressure of the air at any temperature, to the maximum of saturation vapor pressure at the same temperature. Relative humidity in percents is defined as

$$H = 100\frac{P_w}{P_s}, \tag{13.1}$$

where P_w is the partial pressure of water vapor, and P_s is the pressure of saturated water vapor at a given temperature. The value of H expresses the vapor content as a percentage of the concentration required to cause the vapor saturation, that is, the formation of water droplets (dew) at that temperature. An alternative way to present RH is as a ratio of the mole fraction of water vapor in a space to the mole fraction of water vapor in the space at saturation.

The value of P_w together with partial pressure of dry air P_a is equal to pressure in the enclosure, or to the atmospheric pressure P_{atm} if the enclosure is open to the atmosphere:

$$P_{\mathrm{w}} + P_{\mathrm{a}} = P_{atm}. \tag{13.2}$$

At temperatures above the boiling point, water pressure could displace all other gases in the enclosure. The atmosphere would then consist entirely of superheated steam. In this case, $P_w = P_{atm}$. At temperatures above 100°C, RH is a misleading indicator of moisture content because at these temperatures P_s is always more the P_{atm}, and maximum RH never can reach 100%. Thus, at normal atmospheric pressure and temperature of 100°C, the maximum RH is 100%, while at 200°C it is only 6%. Above 374°C, saturation pressures are not thermodynamically specified.

Dewpoint temperature – the temperature at which the partial pressure of the water vapor present would be at its maximum, or saturated vapor condition, with respect to equilibrium with a plain surface of ice. It also is defined as the temperature to which the gas–water vapor mixture must be cooled isobarically (at constant pressures) to induce frost or ice (assuming no prior condensation). The dewpoint is the temperature at which relative humidity is 100%. In other words, the dewpoint is the temperature that the air must reach for the air to hold the maximum amount of moisture it can. When the temperature cools to the dewpoint, the air becomes saturated and fog, or dew or frost can occur.

The following equations calculate the dewpoint from relative humidity and temperature [4]. All temperatures are in Celsius.

The saturation vapor pressure over water is found from

$$EW = 10^{0.66077+7.5\frac{t}{237+t}} \tag{13.3}$$

and the dewpoint temperature is found from the approximation

$$DP = \frac{237.3(0.66077 - \log_{10} EW_{RH})}{\log_{10} EW_{RH} - 8.16077} t \tag{13.4}$$

where

$$EW_{RH} = \frac{EW \cdot RH}{100}$$

Relative humidity displays an inverse relationship with the absolute temperature. Dewpoint temperature is usually measured with a chilled mirror. However, below 0°C dewpoint, the measurement becomes uncertain as moisture eventually freezes and a crystal lattice growth will slowly occur, much like a snowflake. Nevertheless, moisture can exist for prolonged time below 0°C in a liquid phase, depending on variables such as molecular agitation, rate of convection, sample gas temperature, and contaminations.

To calibrate humidity sensors, a reference source of humidity is required. There are several methods of producing a known humidity level. For example, one can generate a dry air (0% humidity) and steaming moist air (100% humidity) and then mix them in a known proportion. Yet, the most popular method is using saturated salt solutions in water. A dish with the saturated solution is placed in a closed box that is tightly sealed from the atmosphere. The solution generates relative humidity in the free space above the dish with good accuracy. The value of the relative humidity depends on the type of salt used (Table 13.1). Relative humidity is very little dependent on temperature, but strongly dependent on temperature special uniformity. For an accuracy of ±2% RH, temperature uniformity inside the box should be better than 0.5°C.

To make a relative or absolute humidity sensor, any physical effect that relates to the concentration of water molecules can be employed. One of the oldest sensors to

Table 13.1 Relative humidity of saturated salt solutions (from Greenspan L, Huang PH, Wahtstone JR, Aoro RM. 808 10 and OIML recommendations)

Temperature (°C)	Lithium chloride solution LiCl, H_2O	Magnesium chloride solution MgCl, $6H_2O$	Magnesium nitrate solution $Mg(NO_3)_2$, $6H_2O$	Sodium chloride solution NaCl, $6H_2O$	Potassium chloride solution K_2SO_4
5	13	33.6 ± 0.3	58	75.7 ± 0.3	98.5 ± 0.9
10	13	33.5 ± 0.2	57	75.7 ± 0.2	98.2 ± 0.8
15	12	33.3 ± 0.2	56	75.6 ± 0.2	97.9 ± 0.6
20	12	33.1 ± 0.2	55	75.5 ± 0.1	97.6 ± 0.5
25	11.3 ± 0.3	32.8 ± 0.3	53	75.3 ± 0.1	97.3 ± 0.5
30	11.3 ± 0.2	32.4 ± 0.1	52	75.1 ± 0.1	97.0 ± 0.4
35	11.3 ± 0.2	32.1 ± 0.1	50	74.9 ± 0.1	96.7 ± 0.4
40	11.2 ± 0.2	31.6 ± 0.1	49	74.7 ± 0.1	96.4 ± 0.4
45	11.2 ± 0.2	31.1 ± 0.1	–	74.5 ± 0.2	96.1 ± 0.4
50	11.1 ± 0.2	30.5 ± 0.1	46	74.6 ± 0.9	95.8 ± 0.5
55	11.0 ± 0.2	29.9 ± 0.2	–	74.5 ± 0.9	–

measure humidity was a hair tension transducer. The hair may be from a human or a animal, its tension is the function of ambient humidity. If the hair is stretched between two anchor points, the tension can be converted to electrical signal by any appropriate force sensor. Tension is stronger at dry air, while the hair relaxes at humid air. On this principle, the folk art "weather houses" were operating: human figure dolls changed positions under the control of the hair tension and, thus "predicting" weather. In the following sections, we examine the most popular modern humidity and moisture sensors.

13.2 Capacitive Sensors

An air-filled capacitor may serve as a relative humidity sensor because moisture in the atmosphere changes air's electrical permitivity according to the following equation [5]:

$$\kappa = 1 + \frac{211}{T}\left(P + \frac{48P_s}{T}H\right)10^{-6}, \tag{13.5}$$

where T is the absolute temperature in Kelvin, P is the pressure of moist air in mmHg, P_s is the pressure of saturated water–vapor at temperature T in mmHg, and H is the relative humidity in %. Equation (13.3) shows that the dielectric constant of moist air, and, therefore, the capacitance, is proportional to the relative humidity.

Instead of air, the space between the capacitor plates can be filled with an appropriate isolator whose dielectric constant changes significantly upon being subjected to humidity. The capacitive sensor may be formed of a hygroscopic

polymer film with metallized electrode s deposited on the opposite sides. In one design [6], the dielectric was composed of a hygrophilic polymer thin film (8–12 μm thick) made of cellulose acetate butyrate and the dimetylephtalate as plasticizer. A size of the film sensor is 12 × 12 mm. The 8-mm-diameter gold porous disk electrodes (200 Å thick) were deposited on the polymer by the vacuum deposition. The film was suspended by a holder, and the electrodes were connected to the terminals. The capacitance of such a sensor is approximately proportional to relative humidity H

$$C_h \approx C_o(1 + \alpha_h H), \tag{13.6}$$

where C_o is the capacitance at $H = 0$.

For the use with capacitive sensors, 2% accuracy in the range from 5 to 90% RH can be achieved with a simple circuit shown in Fig. 13.1. The sensor and the circuit transfer characteristics are shown in Fig. 13.2. The sensor's nominal capacitance at 75% RH is 500 pF. It has a quasilinear transfer function with the offset at zero humidity of about 370 pF and the slope of 1.7 pF/% RH. The circuit effectively performs two functions: makes a capacitance-to-voltage conversion and subtracts the offset capacitance to produce an output voltage with zero intercept. The heart of the circuit is a self-clocking analog switch LT1043 that multiplexes several capacitors at the summing junction (virtual ground) of the operational amplifier U_1. The capacitor C_1 is for the offset capacitance subtraction, while the capacitor C_1 is connected in series with the capacitive sensor S_1. The average voltage across the sensor must be zero; otherwise, electrochemical migration could damage it permanently. Nonpolarized

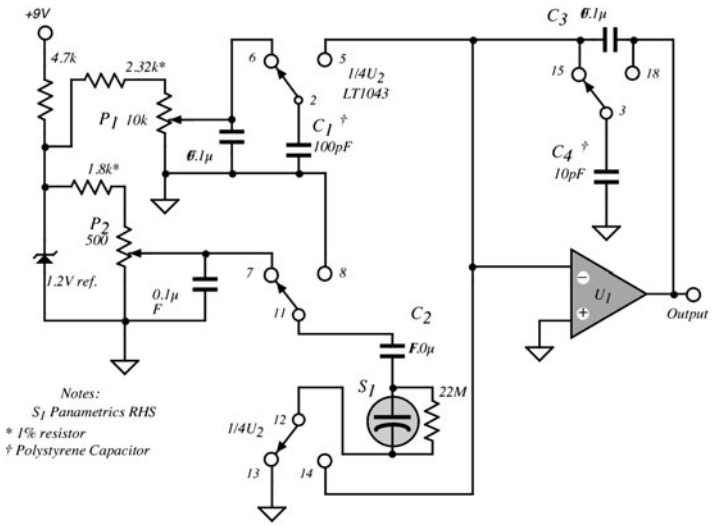

Fig. 13.1 Simplified circuit for measuring humidity with a capacitive sensor (adapted from Sashida and Sakaino [6])

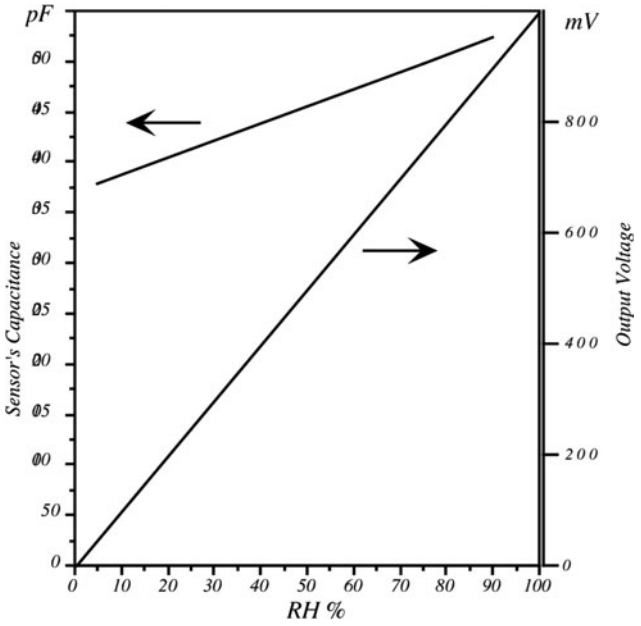

Fig. 13.2 Transfer functions of a capacitive sensor and a system

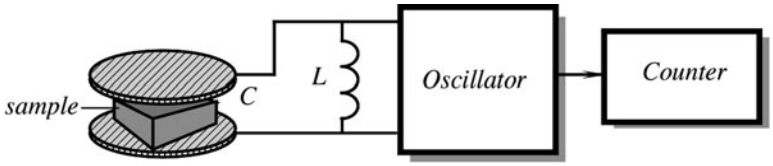

Fig. 13.3 Capacitive moisture sensing system

capacitor C_2 protects the sensor against building up any dc charge. Trimpot P_2 adjusts the amount of charge delivered to the sensor, and P_1 trims the offset charge that is subtracted from the sensor. The net charge is integrated with the help of the feedback capacitor C_3. Capacitor C_4 maintains dc output when the summing junction is disconnected from the sensor.

A similar technique can be used for measuring moisture in material samples [7]. Figure 13.3 shows a block diagram of the capacitive measurement system where the dielectric constant of the sample changes frequency of the oscillator. This method of moisture measurement is quite useful in the process control of the pharmaceutical products. Dielectric constants of most of the medical tablets is quite low (between 2.0 and 5.0) when compared with that of water (Fig. 3.7). The sampled material is placed between two test plates that form a capacitor connected into an *LC*-oscillating circuit. The frequency is measured and related to the moisture. The best way to reduce variations attributed to environmental conditions, such as

temperature and room humidity, is the use of a differential technique. That is, the frequency shift $\Delta f = f_0 - f_1$ is calculated, where f_0 and f_1 are frequencies produced by the empty container and that filled with the sampled material, respectively. The method has some limitations; for instance, its accuracy is poor when measuring moistures below 0.5%, the sample must be clean of foreign particles having relatively high dielectric constants – examples are metal and plastic objects – packing density and fixed sample geometry must be maintained.

A thin film capacitive humidity sensor can be fabricated on a silicon substrate [8]. A layer of SiO_2 3,000 Å thick is grown on an *n*-Si substrate (Fig. 13.4b). Two-metal electrodes are deposited on the SiO_2 layer. They are made of aluminum-, chromium-, or phosphorous-doped polysilicon (LPCVD). The electrode thickness is in the range from 2,000 to 5,000 Å. The electrodes are shaped in an interdigitized pattern as shown in Fig. 13.4a. To provide additional temperature compensation, two temperature sensitive resistors are formed on the same substrate. The top of the sensor is coated with a dielectric layer. For this layer, several materials can be used, such as chemically vapor-deposited SiO_2 or phosphorosilicate glass (CVD PSG). The thickness of the layer is in the range from 300 to 4,000 Å.

A simplified equivalent electrical circuit is shown in Fig. 13.5. Each element of the circuit represents a *RC*-transmission line [9]. When the relative humidity increases, the distributed surface resistance drops, and the equivalent capacitance between the terminals 1 and 2 grows. The capacitance is frequency-dependent; hence, for the low humidity range measurement, frequency should be selected near 100 Hz, while for the higher humidities, it is in the range between 1 and 10 kHz.

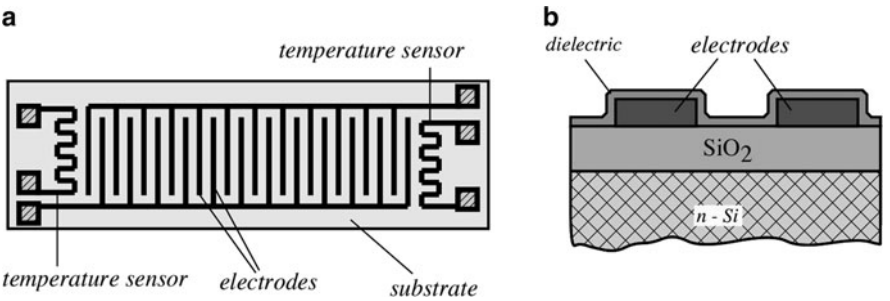

Fig. 13.4 Capacitive thin-film humidity sensor interdigitized electrodes form capacitor plates (**a**); cross-section of the sensor (**b**)

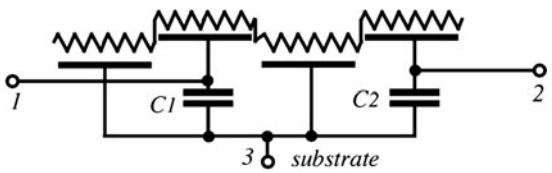

Fig. 13.5 Simplified equivalent electric circuit of a capacitive thin-film humidity sensor

13.3 Electrical Conductivity Sensors

Resistances of many nonmetal conductors generally depend on their water content, as it was discussed in Sect. 3.5.4. This phenomenon is the basis of a resistive humidity sensor or hygristor. A general concept of a conductive hygrometric sensor is shown in Fig. 13.6a. The sensor is fabricated on a ceramic (alumina) substrate. The moisture-sensing material has relatively low resistivity, which changes significantly under varying humidity conditions. The material is deposited on the top of two interdigitized electrodes to provide a large contact area. When water molecules are absorbed by the upper layer, resistivity between the electrodes changes, which can be measured by an electronic circuit. The first such sensor was developed by F.W. Dunmore in 1935 with a hygroscopic film consisting of 2–5% aqueous solution of LiCl [10]. Another example of a conductive humidity sensor is the so-called Pope element, which contains a polystyrene film treated with sulfuric acid to obtain the desired surface-resistivity characteristics.

Other promising materials for the fabrication of a film in a conductivity sensor are solid polyelectrolytes because their electrical conductivity varies with humidity. Long-term stability and repeatability of these compounds, while generally not too great, can be significantly improved by using the interpenetrating polymer networks and carriers, and supporting media. When measured at 1 kHz, an experimental sample of such a film has demonstrated a change in impedance from 10 MΩ to 100 Ω while RH was changing from 0 to 90% [11]. A conductometric humidity sensor can be mounted at the tip of a probe or on a circuit board (Fig. 13.6b).

A solid-state humidity sensor can be fabricated on a silicon substrate (Fig. 13.7b). The silicon must be of a high conductance [12], which provides an electrical path from the aluminum electrode vacuum deposited on its surface. An oxide layer is formed on the top of the conductive aluminum layer, and on the top of that, another electrode is formed. The aluminum layer is anodized in a manner to form a porous oxide surface. The average cross-sectional dimension of pores is sufficient to allow penetration by water molecules. The upper electrode is made in a form of porous gold that is permeable to gas and, at the same time, can provide

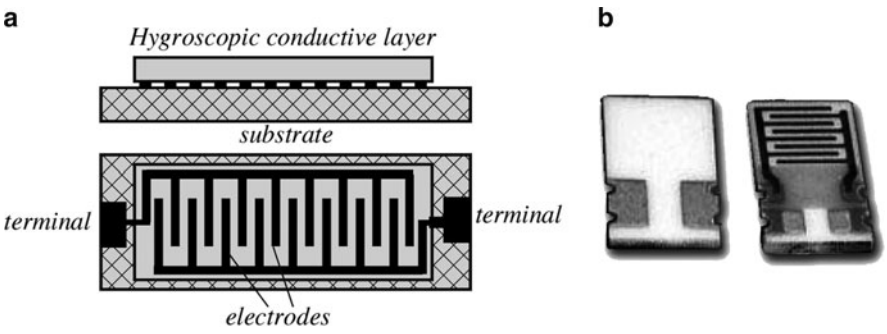

Fig. 13.6 Composition of a conductive humidity sensor (**a**); back and front of a humidity sensor for surface mounting (**b**)

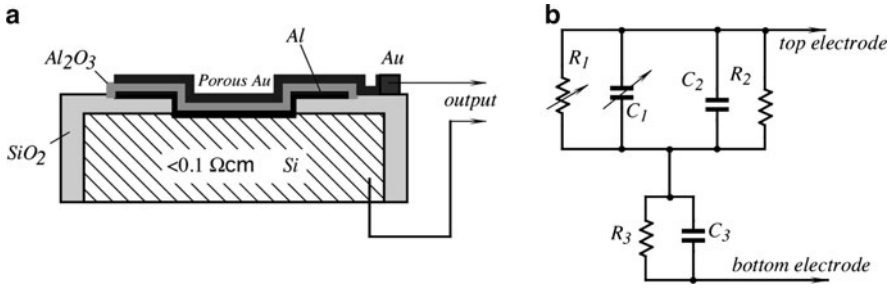

Fig. 13.7 Structure of Al_2O_3 thin film moisture sensor (**a**). Simplified equivalent circuit of the sensor (**b**)

electrical contact. Electrical connections are made to the gold and silicon layers. Aluminum oxide (Al_2O_3), like numerous other materials, readily absorbs water when in contact with a gas mixture containing water in the vapor phase. The amount of sorption is proportional to the water vapor partial pressure and inversely proportional to the absolute temperature. Aluminum oxide is a dielectric material. Its dielectric constant and surface resistivity are modified by the physisorption of water. For this reason, this material lends itself as a humidity-sensing compound.

Figure 13.7b shows an electrical equivalent circuit of the sensor [13]. The values of R_1 and C_1 depend on the Al_2O_3 average pore sizes and density. These components of resistance and capacitance vary with the number of water molecules that penetrate the pores and adhere to the surface. R_2 and C_2, respectively, represent the resistance and capacitance components of the bulk oxide material between the pores and are therefore unaffected by moisture. C_3 is an equivalent series capacitance term as determined by the measurement of the total resistance components in a dry atmosphere at very low frequencies. The sensor's resistance becomes very large ($>10^8$ **Ω**) as the frequency approaches dc. Thus, the measurement of humidity involves the measurement of the sensor's impedance. The residual of nonhumidity-dependent resistance and capacitance terms that exist in a typical sensor shunt the humidity-dependent variables, thus causing the continuous reduction in slope (sensitivity) as the humidity is lowered, which, in turn, reduces the accuracy at lower humidities. Since temperature is a factor in humidity measurement, the sensor usually combines in the same package a humidity sensor, a thermistor, and a reference capacitance, which is protected against humidity influence and has a low-temperature coefficient.

R_1 and C_1 are moisture-dependent variable terms, R_2 and C_2 are shunting terms of bulk oxide between pores (unaffected by moisture), R_3 and C_3 are series terms below pores (unaffected by moisture).

In agriculture and geology, investigation of soils is a serious business. Many soil moisture monitors operate on the principle of conductivity measurement [14]. Pure water is not a good conductor of electricity. Because the electrical current is transported by ions in a solution, the conductivity increases as the concentration of ions increases. Thus, conductivity increases as water dissolved ionic species.

Typical conductivity of waters[2]:

Deionized water	5.5×10^{-6} S/m
Drinking water	0.005–0.05 S/m
Sea water	5 S/m

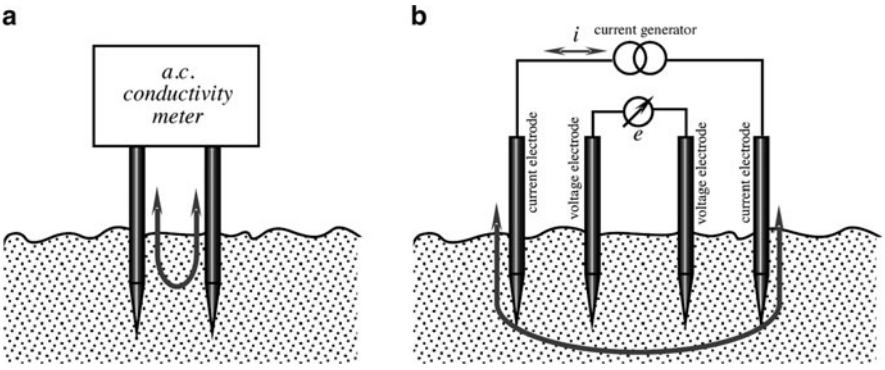

Fig. 13.8 Measurement of soil electrical conductivity with two-electrode (**a**) and four-electrode (**b**) systems

Soil is composed of a great variety of minerals and organic materials existing in solid, gaseous, and aqueous states. The aqueous components, that is water dissolved, are the main contributors to the soil electrical conductivity. Water in soils, depending on the soil composition, ranges in conductivity from 0.01 to 8 S/m and is the main contributor to soil electrical conductivity. To monitor the soul water content, several methods are currently employed. One uses a simple two-electrode probe as shown in Fig. 13.8a where electrical resistivity is measured by monitoring the voltage and current flow between the electrodes inserted into the soil sample. Measurement should be made with ac current to prevent the electrode polarization. An improved method is based on the four-electrode resistance measurement system (Fig. 13.8b) whose principle is described in Sect. 5.12.2. Advantage of this system is the absence of electrical current in the voltage electrodes, and thus a better accuracy. The current electrodes inject into the soil ac current i from a high output impedance current generator, while the voltage electrodes are connected to a high input impedance amplifier to measure the ac voltage e. That voltage depends only on the soil conductivity between the voltage electrodes and, thanks to a high output impedance of the current generator, is independent of the soil properties between the current electrodes.

[2]Conductivity is measured in units of siemens (S), which is a reciprocal function of specific resistivity. See Sect. 3.5.1 in Chap. 3.

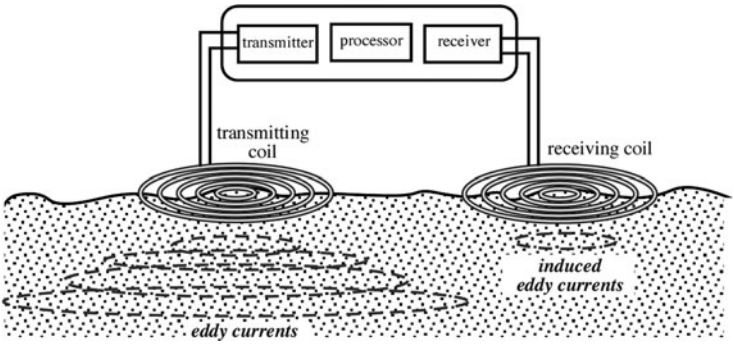

Fig. 13.9 Operating principle of electromagnetic conductivity measurement of soil

Sometimes, measurement should be made at the appreciable depth where using contact electrodes is impractical. For such cases, the electromagnetic conductivity measurement is performed with two coils – transmitting and receiving (Fig. 13.9). A transmitting coil is located at one end of the instrument. It induces circular eddy-current loops into the soil. The magnitude of these loops is directly proportional to the electrical conductivity of the soil in the vicinity of that loop. Each current loop generates a secondary electromagnetic field that is proportional to the value of the current flowing within the loop. A fraction of the secondary induced electromagnetic field from each eddy loop is intercepted by the receiver coil of the instrument, and the sum of these induced signals is amplified and detected as output voltage, which relates to a depth-weighted bulk soil electrical conductivity. The receiver coil measures amplitudes and phases of the secondary magnetic field. The amplitude and phase of the secondary field will differ from those of the primary field as a result of soil properties (e.g., clay content, water content, and salinity), spacing of the coils, and their distance from the soil surface [15].

13.4 Thermal Conductivity Sensor

Using thermal conductivity of gas to measure humidity can be accomplished by a thermistor-based sensor (Fig. 13.8a) [16]. Two tiny thermistors (R_{t1} and R_{t2}) are supported by thin wires to minimize thermal conductivity loss to the housing. The left thermistor is exposed to the outside gas through small venting holes, while the right thermistor is hermetically sealed in dry air. Both thermistors are connected into a bridge circuit (R_1 and R_2), which is powered by voltage $+E$. The thermistors develop self-heating due to the passage of electric current. Their temperatures rise up to 170°C over the ambient temperature. Initially, the bridge is balanced in dry air to establish a zero reference point. The output of this sensor gradually increases as absolute humidity rises from zero. At about 150 g/m^3, it reaches the saturation and then decreases with a polarity change at about 345 g/m^3 (Fig. 13.8b). In this

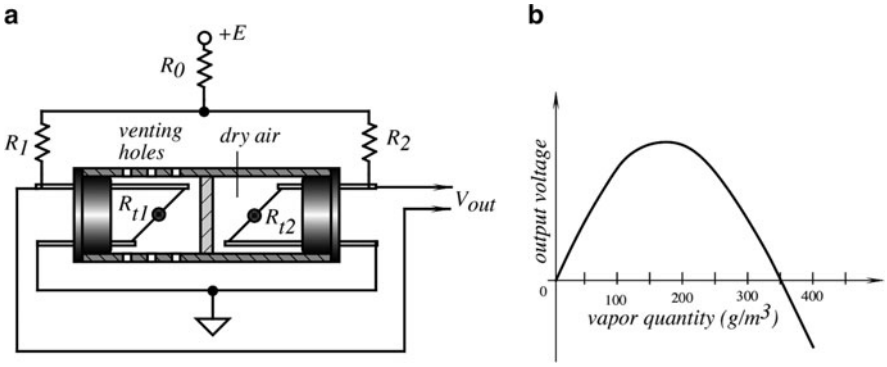

Fig. 13.10 Absolute humidity sensor with self-heating thermistors: (**a**) design and electrical connection and (**b**) output voltage

device, it is important to conduct measurement in a near-still air with very little air flow through the venting holes. Otherwise, air convection will cause additional cooling and thus cause errors in measurement (Fig. 13.10).

13.5 Optical Hygrometer

Most of the humidity sensors exhibit some repeatability problems, especially hysteresis with a typical value from 0.5 to 1% RH. In precision process control, this may be a limiting factor; therefore, indirect methods of humidity measurements should be considered. The most efficient method is a calculation of absolute or relative humidity through dewpoint temperature. As was indicated earlier in this chapter, the dewpoint is the temperature at which liquid and vapor phases of water (or any fluid for that matter) are in equilibrium. The temperature at which the vapor and solid phases are in equilibrium is called frostpoint. At the dewpoint, only one value of saturation vapor pressure exists. Hence, absolute humidity can be measured from this temperature as long as the pressure is known. The optimum method of moisture measurement by which the minimum hysteresis effects are realized requires the use of optical hygrometry. The cost of an optical hygrometer is considerably greater, but if the benefit of tracking low-level moisture enhances product yield and quality, the cost is easily justified.

The basic idea behind the optical hygrometer is the use of a mirror whose surface temperature is precisely regulated by a thermoelectric heat pump. The mirror temperature is controlled at a threshold of the formation of dew. Sampled air is pumped over the mirror surface and, if the mirror temperature crosses a dewpoint, releases moisture in the form of water droplets. The reflective properties of the mirror change at water condensation because water droplets scatter light rays. This can be detected by an appropriate photodetector. Figure 13.11 shows a

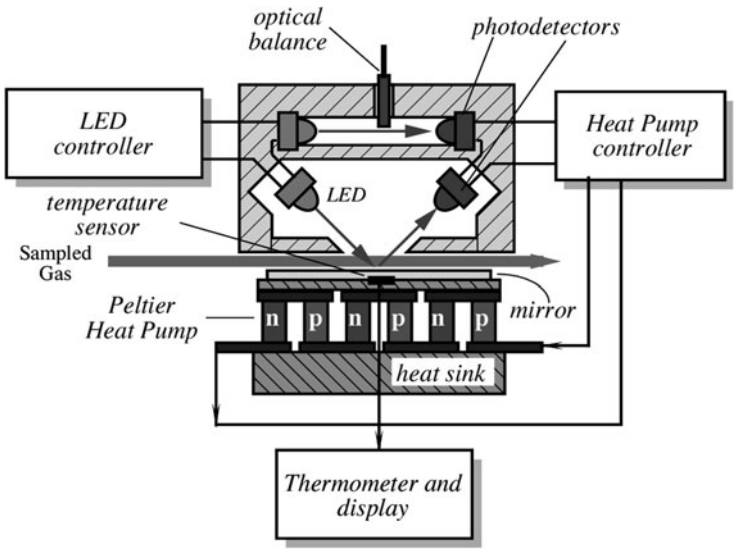

Fig. 13.11 Chilled mirror dewpoint sensor with an optical bridge

simplified block diagram of a chilled mirror hygrometer. It is comprised of a heat pump operating on a Peltier effect. The pump removes heat from a thin mirrored surface that has an imbedded temperature sensor. That sensor is part of a digital thermometer that displays the temperature of the mirror. The hygrometer's circuit is of a differential type, where the top optocoupler, a light-emitting diode (LED), and a photodetector are used for the compensation of drifts, while the bottom optocoupler is for measuring the mirror reflectivity. The sensor's symmetry can be balanced by a wedged optical balance inserted into the light path of the upper optocoupler. The lower optocoupler is positioned at 45° angle with respect to the mirror. Above the dewpoint, the mirror is dry and its reflectivity is the highest. The controller lowers the temperature of the mirror through the heat pump. At the moment of the water condensation, the mirror reflectivity drops abruptly, which caused reduction in a photocurrent in the photodetector. The photodetector signals pass to the controller to regulate electric current through the heat pump to maintain its surface temperature at the level of a dewpoint, where no additional condensation or evaporation from the mirror surface occurs. Actually, water molecules are continuously being trapped and are escaping from the surface, but the average net level of the condensate density does not change once equilibrium is established.

Since the sensed temperature of the mirrored surface precisely determines the actual prevailing dewpoint, this is considered the moisture's most fundamental and accurate method of measurement. Hysteresis is virtually eliminated, and sensitivity is near 0.03°C DP (dewpoint). From the dewpoint, all moisture parameters such as % RH and vapor pressure are obtainable once the prevailing temperature and pressure are known.

There are several problems associated with the method. One is a relatively high cost, the other is a potential mirror contamination, and the third is relatively high power consumption by the heat pump. Contamination problems can be virtually eliminated with the use of particle filters and a special technique that deliberately cools the mirror well below the dewpoint to cause excessive condensation, with the following fast rewarming. This flashes the contaminants keeping the mirror clean [17].

13.6 Oscillating Hygrometer

The idea behind the oscillating hygrometer is similar to that behind the optical-chilled mirror sensor. The difference is that the measurement of the dewpoint is made not by the optical reflectivity of the surface, but rather by detecting the changing mass of the chilled plate. The chilled plate is fabricated of a thin quartz crystal, which is a part of an oscillating circuit. This implies the other name for the sensor: the piezoelectric hygrometer, because the quartz plate oscillation is based on the piezoelectric effect. A quartz crystal is thermally coupled to the Peltier cooler (see Sect. 3.9.2), which controls the temperature of the crystal with a high degree of accuracy (Fig. 13.12). When the temperature drops to that of a dewpoint, a film of water vapor deposits on the exposed surface of the quartz crystal. Since the mass of the crystal changes, the resonant frequency of the oscillator shifts from f_0 to f_1. The new frequency f_1 corresponds to a given thickness of the water layer. The frequency shift controls current through the Peltier cooler, thus changing the temperature of the quartz crystal to stabilize at the dewpoint temperature. The major difficulty in designing the piezoelectric hygrometer is in providing an adequate thermal coupling between the cooler and the crystal, while maintaining small size of the crystal at a minimum mechanical loading [18]. Naturally, this method may be employed by using the SAW sensors, similar to that of Fig. 12.9 where the stimulus place is the area subjected to the sampled gas.

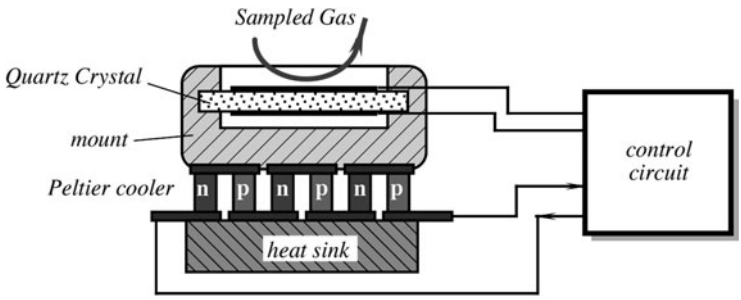

Fig. 13.12 Oscillating humidity sensor

References

1. Quinn FC (1985) The most common problem of moisture/humidity measurement and control. In: Moisture and humidity. Proceedings of the 1985 International Symposium on Moisture and Humidity, ©ISA, Washington, DC, pp. 1–5
2. Carter EF (ed) (1966) Dictionary of inventions and discoveries. Crane, Russak, New York, F. Muller
3. Baughman EH, Mayes D (1989) NIR applications to process analysis. Am Lab 21 (10):54–58
4. Berry FA Jr. (1945) Handbook of meteorology. McGraw-Hill, New York, p 343
5. Conditioner circuit, AN3 (1990) Linear Technology, Applications Handbook
6. Sashida T, Sakaino Y (1985) An interchangeable humidity sensor for an industrial hygrometer. In: Moisture and humidity. Proceedings of the International Symposium on Moisture and Humidity, Washington, DC, April 15–18
7. Carr-Brion K (1986) Moisture sensors in process control. Elsevier, New York
8. Jachowicz, RS, Dumania P (1985) Evaluation of thin-film humidity sensor type MCP-MOS. In: Moisture and humidity. Proceedings of the International Symposium on Moisture and Humidity, Washington, DC, April 15–18
9. Jachowicz RS, Senturia SD (1982) A thin film humidity sensor. Sens Actuators 2:1981
10. Norton HN (1989) Handbook of transducers. Prentice Hall, Englewood Cliffs, NJ
11. Sakai Y, Sadaoka Y, Matsuguchi M, Hirayama K (1991) Water resistive humidity sensor composed of interpenetrating polymer networks of hydrophilic and hydrophobic methacrylate. In: Transducers'91. International Conference on Solid-state Sensors and Actuators. Digest of Technical Papers, pp 562–565, IEEE
12. Fong V (1985) Al_2O_3 moisture sensor chip for inclusion in microcircuit package and the new MIL standard for moisture content. In: Moisture and humidity. Proceedings of the 1985 International Symposium on Moisture and Humidity. Washington, DC, pp 345–357, ISA
13. Harding JC Jr (1985) Overcoming limitations inherent to aluminum oxide humidity sensors. In: Moisture and humidity. Proceedings of the 1985 International Symposium on Moisture and Humidity. Washington, DC, pp 367–378, ISA
14. Hilhorst MA (2000) A pore water conductivity sensor. Soil Sci Soc Am J 64:1922–1925
15. Hendrickx JMH, Kachanoski RG (2002) Indirect measurement of solute concentration: Nonintrusive electromagnetic induction. In: Dane JH, Topp GC (eds) Methods of soil analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI, pp 1297–1306
16. Miura T (1985) Thermistor humidity sensor for absolute humidity measurements and their applications. In: Moisture and humidity. Proceedings of the International Symposium on Moisture and Humidity, Washington, DC, April 15–18
17. Harding JC Jr. (1985) A chilled mirror dewpoint sensor/psychrometric transmitter for energy monitoring and control systems. In: Moisture and humidity. Proceedings of the International Symposium on Moisture and Humidity, Washington, DC, April 15–18
18. Porlier C (1991) Chilled piezoelectric hygrometer: sensor interface design. In: Sensors Expo Proceedings, 107B-7, Helmers Publishing, Dublin, NH, U.S.A.

.

Chapter 14
Light Detectors

There is nothing more practical than a good theory.

– Kurt Lewin

14.1 Introduction

Detectors of electromagnetic radiation in the spectral range from ultraviolet to far infrared are called light detectors. From the standpoint of a sensor designer, absorption of photons by a sensing material may result in either a quantum or thermal response. Therefore, all light detectors are divided into two major groups that are called quantum and thermal. The quantum detectors operate from the ultraviolet to mid-infrared spectral ranges, while thermal detectors are most useful in the mid- and far-infrared spectral ranges where their efficiency at room temperatures exceeds that of the quantum detectors. In this chapter, we cover both types. For description of highly sensitive photon sensors called photomultipliers refer to Sect. 15.1.

Solid-state quantum detectors (photovoltaic and photoconductive devices) relay on the interaction of individual photons with a crystalline lattice of semiconductor materials. Their operations are based on the photoeffect that was discovered by Albert Einstein, which won him the Nobel Prize. In 1905, he made a remarkable assumption about the nature of light that at least under certain circumstances, its energy was concentrated into localized bundles, later named photons. The energy of a single photon is given by

$$E = h\nu, \tag{14.1}$$

where ν is the frequency of light, and $h = 6.626075 \times 10^{-34}$ Js (or 4.13567×10^{-15} eVs) is Planck's constant derived on the basis of the wave theory of light. When a photon strikes a surface of a conductor, it may result in the generation of a

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_14, © Springer Science+Business Media, LLC 2010

free electron. Part (ϕ) of the photon energy E is used to detach the electron from the surface, while the other part gives to the electron its kinetic energy. The photoelectric effect can be described as

$$h\nu = \phi + K_m, \tag{14.2}$$

where ϕ is called the work function of the emitting surface, and K_m is the maximum kinetic energy of the electron upon its exiting the surface. The similar processes occur when a semiconductor pn-junction is subjected to radiant energy: The photon transfers its energy to an electron and, if the energy is sufficiently high, the electron may become mobile, which results in an electric current. If energy is not sufficient for liberating an electron, the photon energy is just converted to heat.

The periodic lattice of crystalline materials establishes the allowed energy bands for electrons that exist within that solid. The energy of any electron within the pure material must be confined to one of these energy bands that may be separated by gaps or ranges of forbidden energies. That is, the electron can have only "permitted" energies.

If light of a proper wavelength [sufficiently high energy of photons, see (14.1)] strikes a semiconductor crystal, the concentration of charge carriers (electrons and holes) in the crystal increases, which manifests in the increased conductivity of a crystal

$$\sigma = e(\mu_e n + \mu_h p), \tag{14.3}$$

where e is the electron charge, μ_e is the electron mobility, μ_h is the hole mobility, and n and p are the respective concentrations of electrons and holes.

Figure 14.1a shows energy bands of a semiconductor material, where E_g is the magnitude in eV of the forbidden band gap. The lower band is called the valence band, which corresponds to those electrons that are bound to specific lattice sites within the crystal. In the case of silicon or germanium, they are parts of the covalent bonding that constitute the interatomic forces within the crystal. The next higher-lying band is called the conduction band and represents electrons that are free to migrate through the crystal. Electrons in this band contribute to the electrical conductivity of the material. The two bands are separated by the band gap, the size

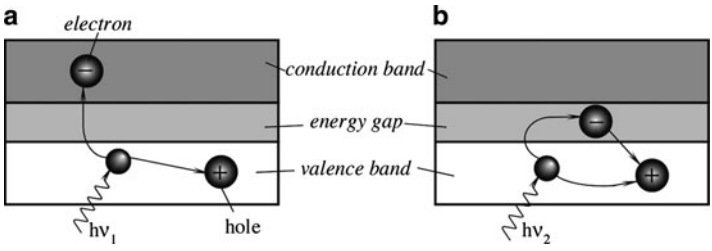

Fig. 14.1 Photoeffect in a semiconductor for high (**a**) and low (**b**) energy photons

Table 14.1 Band gaps and longest wavelengths for various semiconductors (after [1])

Material	Band gap (eV)	Longest wavelength (μm)
ZnS	3.6	0.345
CdS	2.41	0.52
CdSe	1.8	0.69
CdTe	1.5	0.83
Si	1.12	1.10
Ge	0.67	1.85
PbS	0.37	3.35
InAs	0.35	3.54
Te	0.33	3.75
PbTe	0.3	4.13
PbSe	0.27	4.58
InSb	0.18	6.90

of which determines the whether the material is classified as a semiconductor or an isolator. The number of electrons within the crystal is just adequate to completely fill all available sites within the valence band. In the absence of thermal excitation, both isolators and semiconductors would therefore have a configuration in which the valence band is completely full, and the conduction band is completely empty. Under these imaginable circumstances, neither would theoretically show any electrical conductivity.

In a metal, the highest occupied energy band is not completely full. Therefore, electrons can easily migrate throughout the material. Metals are characterized by very high electrical conductivity. In isolators or semiconductors, on the other hand, the electron must first cross the energy band gap in order to reach the conduction band, and the conductivity is therefore many orders of magnitude lower. For isolators, the band gap is usually 5 eV or more, whereas for semiconductors, the gap is considerably less (Table 14.1). Note that the longest the wavelength (lower frequency of a photon), the less energy is required to originate a photoeffect.

When the photon of frequency v_1 strikes the crystal, its energy is high enough to separate the electron from its site in the valence band and push it through the band gap into a conduction band at a higher energy level. In that band, the electron is free to serve as a current carrier. The deficiency of an electron in the valence band creates a hole that also serves as a current carrier. This is manifested in the reduction of specific resistivity of the material. On the other hand, Fig. 14.1b shows that a photon of lower frequency v_2 does not have sufficient energy to push the electron through the band gap. The energy is released as heat without creating current carriers.

The energy gap serves as a threshold below which the material is not light sensitive. However, the threshold is not abrupt. Throughout the photon-excitation process, the law of conservation of momentum applies. The momentum and density of hole-electron sites are higher at the center of both the valence and conduction bands, and fall to zero at the upper and lower ends of the bands. Therefore, the probability of an excited valence-band electron finding a site of like momentum in the conduction band is greater at the center of the bands, and the lowest at the ends of the bands. Therefore, the response of a material to photon energy increases from E_g

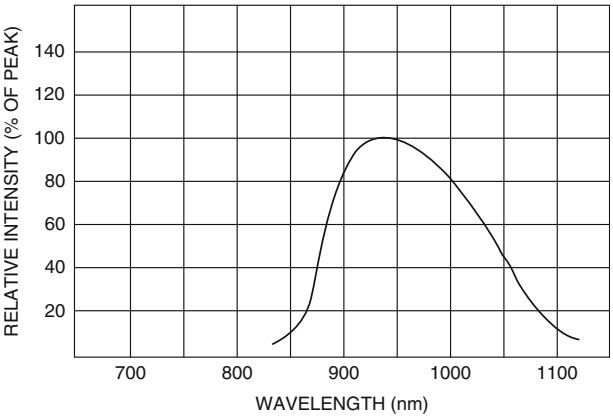

Fig. 14.2 Spectral response of an infrared photodiode

gradually to its maximum and then falls back to zero at the energy corresponding to the difference between the bottom of the valence band and the top of the conduction band. A typical spectral response of a semiconductive material is shown in Fig. 14.2. The light response of a bulk material can be altered by adding various impurities. They can be used to reshape and shift a spectral response of the material. All devices that directly convert photons of electromagnetic radiation into charge carriers are called quantum detectors, which are generally produced in the form of photodiodes, phototransistors, and photoresistors.

When comparing the characteristics of different photodetectors, the following specifications usually should be considered:

Noise equivalent power (*NEP*) is the amount of light equivalent to the intrinsic noise level of the detector. Stated differently, it is the light level required to obtain a signal-to-noise ratio equal to unity. Since the noise level is proportional to the square root of the bandwidth, the NEP is expressed in units of W/Hz^{-2}

$$NEP = \frac{\text{noisecurrent}(A/\sqrt{Hz})}{\text{radiantsensitivity at } \lambda_p (A/W)} \tag{14.4}$$

D^* (*D-star*) refers to the detectivity of a detector's sensitive area of 1 cm^2 and a noise bandwidth of 1 Hz

$$D^* = \frac{\sqrt{\text{area}(cm^2)}}{NEP} \tag{14.5}$$

Detectivity is another way to measure the sensor's signal-to-noise ratio. Detectivity is not uniform over the spectral range for operating frequencies; therefore, the

chopping frequency and the spectral content must be also specified. The detectivity is expressed in the units of Hz^{-2}/W. It can be said that the higher the value of D^*, the better the detector.

IR cutoff wavelength (λ_c) represents the long-wavelength limit of spectral response, and often is listed as the wavelength at which the detectivity drops by 10% of the peak value.

Maximum current is specified for photoconductive detectors (such as HgCdTe) that operate at constant currents. The operating current never should exceed the maximum limit.

Maximum reverse voltage is specified for Ge and Si photodiodes and photoconductive cells. Exceeding this voltage can cause the breakdown and severe deterioration of the sensor's performance.

Radiant responsivity is the ratio of the output photocurrent (or output voltage) divided by the incident radiant power at a given wavelength, expressed in A/W or V/W.

Dark current I_D for photodiodes is a leakage current at a reverse voltage when the diode is in complete darkness. This current generally is temperature dependent and may vary from pA to μA. It approximately doubles for every 10°C increase in temperature.

Field of view (FOV) is the angular measure of the volume of space where the sensor can respond to the source of radiation.

Junction capacitance (C_j) is similar to the capacitance of a parallel plate capacitor. It should be considered whenever a high-speed response is required. The value of C_j drops with reverse bias and is higher for the larger diode areas.

14.2 Photodiodes

Photodiodes are semiconducting optical sensors, which if broadly defined may even include solar batteries. However, here we consider only the information aspect of these devices rather than the power conversion. In a simple way, the operation of a photodiode can be described as follows.

If a pn-junction is forward biased (positive side of a battery is connected to the p side) and is exposed to light of proper frequency, the current increase will be very small with respect to a dark current. In other words, the bias current is much greater than the current generated by light, and the diode is just a diode, not really useful for sensing light.

If the junction is reverse biased (Fig. 14.3), when light strikes the semiconductor, the current will increase quite noticeably. Impinging photons create electron–hole pairs on both sides of the junction. When electrons enter the conduction band, they start flowing toward the positive side of the battery. Correspondingly, the created holes flow to the negative terminal, meaning that photocurrent i_p flows in the network. Under dark conditions, dark current i_0 is independent of applied voltage and mainly is the result of thermal generation of charge carriers. Thus, a

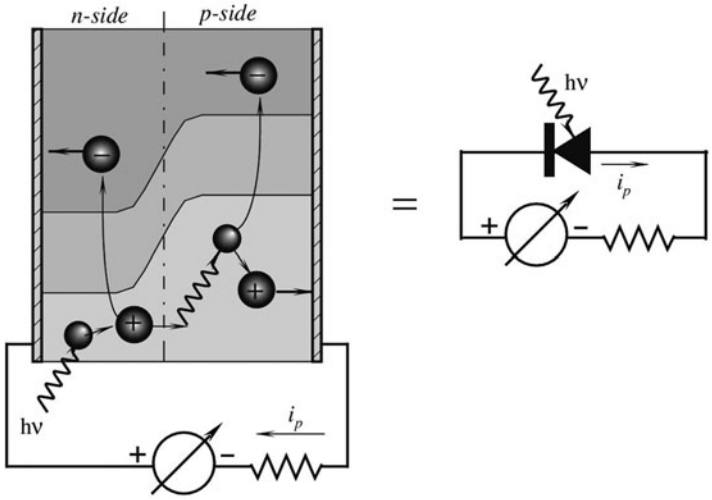

Fig. 14.3 Structure of a photodiode

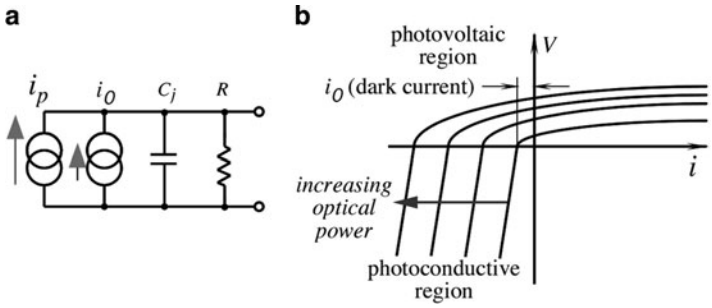

Fig. 14.4 An equivalent circuit of a photodiode (**a**) and its volt-ampere characteristic (**b**)

reverse-biased photodiode electrical equivalent circuit contains two current sources and a *RC* network (Fig. 14.4a).

The process of optical detection involves the direct conversion of optical energy (in the form of photons) into an electrical signal (in the form of electrons). If the probability that a photon of energy hv will produce an electron in a detection is η then the average rate of the production of electrons $\langle r \rangle$ for an incident beam of optical power P is given by [2]

$$\langle r \rangle = \frac{\eta P}{hv} \tag{14.6}$$

The production of electrons due to the incident photons at constant rate $<r>$ is randomly distributed in time and obeys Poisson statistics, so that the probability of the production of m electrons in some measurement interval τ is given by

$$p(m,\tau) = (\langle r\rangle\tau)^m \frac{1}{m!} \mathrm{e}^{-\langle r\rangle\tau} \tag{14.7}$$

The statistics involved with optical detection are very important in the determination of minimum detectable signal levels, and hence the ultimate sensitivity of the sensors. At this point, however, we just note that the electrical current is proportional to the optical power incident on the detector:

$$i = \langle r\rangle e = \frac{\eta e P}{h\nu}, \tag{14.8}$$

where e is the charge of an electron. A change in input power ΔP (due to intensity modulation in a sensor, for instance) results in the output current Δi. Since power is proportional to squared current, the detector's electrical power output varies quadratically with input optical power, making it a "square-law" optical power detector.

The voltage-to-current response of a typical photodiode is shown in Fig. 14.4b. If we attach a high-input-impedance voltmeter to the diode (corresponds to the case when $i = 0$), we will observe that with increasing optical power, the voltage changes in a quite nonlinear fashion. In fact, variations are logarithmic. However, for the short circuit conditions ($V = 0$), that is, when the diode is connected to a current-to-voltage converter (such as in Fig. 5.12b), current varies linearly with the optical power. The current-to-voltage response of the photodiode is given by [3]

$$i = i_0\left(\mathrm{e}^{eV/k_bT} - 1\right) - i_s, \tag{14.9}$$

where i_0 is a reverse "dark current," which is attributed to the thermal generation of electron–hole pairs, i_s is the current due to the detected optical signal, k_b is Boltzmann constant, and T is the absolute temperature. Combining (14.8) and (14.9) yields

$$i = i_0\left(\mathrm{e}^{eV/k_bT} - 1\right) - \frac{\eta e P}{h\nu}, \tag{14.10}$$

which is the overall characteristic of a photodiode. An efficiency of the direct conversion of optical power into electric power is quite low. Typically, it is in the range of 5–10%; however, it is reported that some experimental photocells were able to reach efficiency as high as 90%. In the sensor technologies, however, the photocells are generally not used.

There are two general operating modes for a photodiode: the photoconductive (PC) and the photovoltaic (PV). No bias voltage is applied for the photovoltaic mode. The result is that there is no dark current, so there is only thermal noise present. This allows much better sensitivities at low-light levels. However, the speed response is worst due to an increase in C_j and responsively to longer wavelengths is also reduced.

Figure 14.5a shows a photodiode connected in a PV mode. In this connection, the diode operates as a current-generating device, which is represented in the

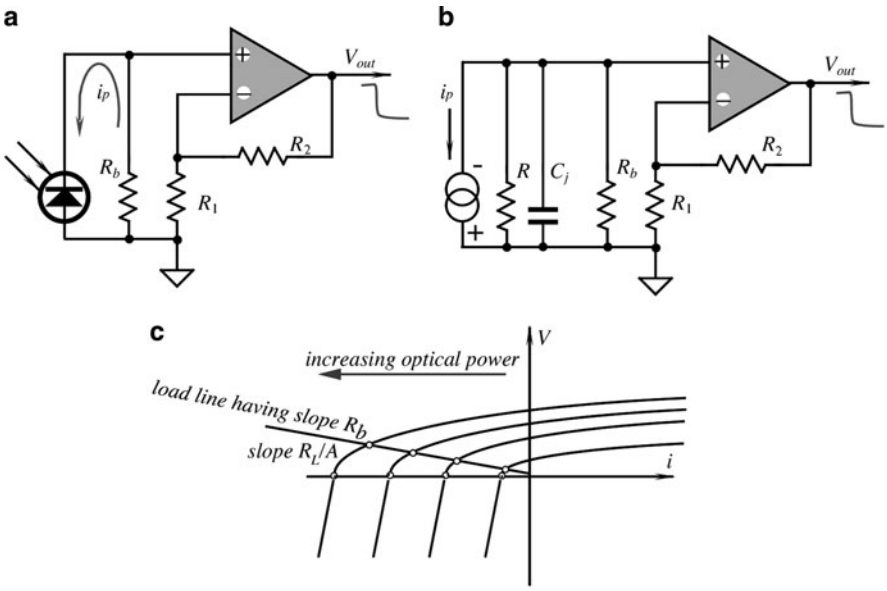

Fig. 14.5 Connection of a photodiode in a photovoltaic mode to a noninverting amplifier (**a**); equivalent circuit (**b**); and loading characteristic (**c**)

equivalent circuit by a current source i_p in Fig. 14.5b. The load resistor R_b determines the voltage developed at the input of the amplifier, and the slope of the load characteristic is proportional to that resistor (Fig. 14.5c).

When using a photodiode in a photovoltaic mode, its large capacitance C_j may limit the speed response of the circuit. During the operation with a direct resistive load, as in Fig. 14.5a, a photodiode exhibits a bandwidth limited mainly by its internal capacitance C_j. The equivalent circuit of Fig. 14.5b models such a bandwidth limit. The photodiode acts primarily as a current source. A large resistance R and the diode capacitance shunt the source. The capacitance ranges from 2 to 20,000 pF depending for the most part on the diode area. In parallel with the shunt is the amplifier's input capacitance (not shown) that results in a combined input capacitance C. The net input network determines the input circuit response rolloff.

To avoid the effect of input capacitance, it is desirable to develop input voltage across the resistor and prevent it from charging the capacitances. This can be achieved by employing a current-to-voltage amplifier (I/V) as shown in Fig. 14.6a. The amplifier and its feedback resistor R_L translate the diode current into a buffered output voltage with excellent linearity. Added to the figure is a feedback capacitor C_L that provides a phase compensation. An ideal amplifier holds its two inputs at the same voltage (ground in the figure), thus the inverting input that is not directly connected to ground is called a virtual ground. The photodiode operates at zero voltage across its terminals, which improves the response linearity

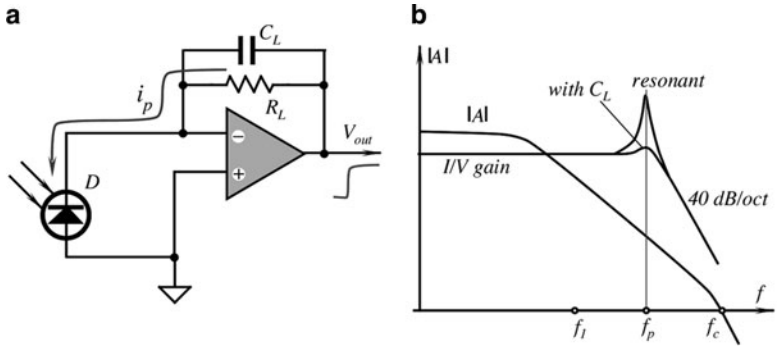

Fig. 14.6 Use of current-to-voltage converter (**a**) and the frequency characteristics (**b**)

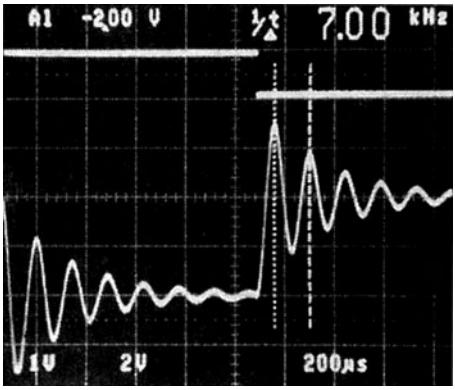

Fig. 14.7 Response of a photodiode with uncompensated circuit (courtesy of Hamamatsu Photonics K.K.)

and prevents charging the diode capacitance. This is illustrated in Fig. 14.7c where the load line virtually is parallel to the current axis, because the line's slope is inversely proportional to the amplifier's open-loop gain A.

In practice, the amplifier's high but finite open-loop gain A limits the performance by developing small albeit nonzero voltage across the diode. Then, the break frequency is defined as

$$f_p = \frac{A}{2\pi R_L C} \approx A f_1 \tag{14.11}$$

where A is the open-loop gain of the amplifier. Therefore, the break frequency is increased by a factor A when compared with f_1. It should be noted that when frequency increases, gain A declines, and the virtual load attached to the photodiode appears to be inductive. This results from the phase shift of gain A. Over most of the amplifier's useful frequency range, A has a phase lag of 90°. The 180° phase inversion by the amplifier converts this to a 90° phase lead that is specific for the inductive impedance. This inductive load resonates with the capacitance of the input

circuit at a frequency equal to f_p (Fig. 14.6b) and may result in an oscillating response (Fig. 14.7) or the circuit instability. To restore stability, a compensating capacitor C_L is placed across the feedback resistor. Value of the capacitor can be found from

$$C_L = \frac{1}{2\pi R_L f_p} = \sqrt{CC_c} \tag{14.12}$$

where $C_c = 1/(2\pi R_L f_c)$ and f_c is the unity-gain crossover frequency of the operational amplifier . The capacitor boosts the signal at the inverting input by shunting R_L at higher frequencies.

When using photodiodes for the detection of low-level light, noise floor should be seriously considered. There are two main components of noise in a photodiode: shot noise and Johnson noise (see Sect. 5.13.1). Besides the sensor, the amplifier's and auxiliary component noise should also be accounted for [see (5.52)].

For the photoconductive operating mode (PC), a reverse bias voltage is applied to the photodiode. The result is a wider depletion region, lower junction capacitance C_j, lower series resistance, shorter rise time, and linear response in photocurrent over a wider range of light intensities. However, as the reverse bias is increased, shot noise increases as well due to increase in a dark current. The PC mode circuit diagram is shown in Fig. 14.8a, and the diode's load characteristic is in Fig. 14.8b. The reverse bias moves the load line into the third quadrant where the response linearity is better than that for the PV mode (the second quadrant). The load lines cross the voltage axis at the point corresponding to the bias voltage E, while the slope is inversely proportional to the amplifier's open-loop gain A. The PC mode offers bandwidths to hundreds of MHz, providing an accompanying increase in the signal-to-noise ratio.

Presently, photodiodes together with interface electronic circuits are available in integral forms and known as light-to-voltage converters. Such an integrated circuit is comprised of a photodiode and current-to-voltage converter as shown in

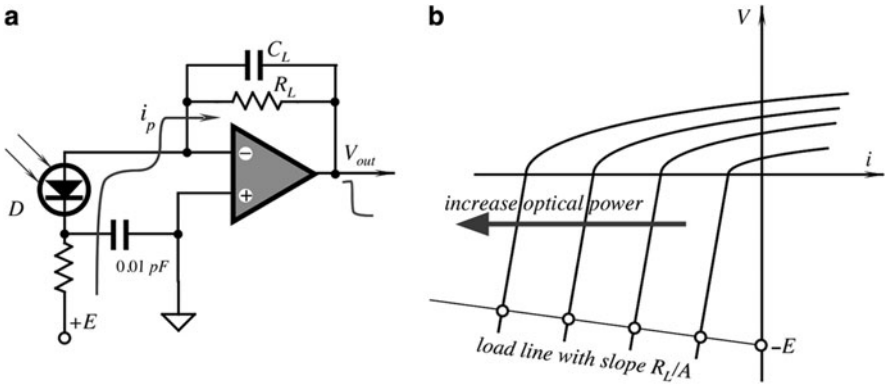

Fig. 14.8 Photoconductive operating mode. Circuit diagram (**a**) and load characteristic (**b**)

Fig. 14.6. An example is a device TSL257T from TAOS (www.taosinc.com), which operates primarily in the visible spectral range from 400 to 800 nm.

14.3 Phototransistor

A photodiode directly converts photons into charge carriers, specifically one electron and one hole (hole–electron pair) per a photon. The phototransistors can do the same, and in addition to provide current gain, resulting in a much higher sensitivity. The collector–base junction is a reverse-biased diode that functions as described earlier. If the transistor is connected into a circuit containing a battery, a photoinduced current flows through the loop, which includes the base–emitter region. This current is amplified by the transistor in the same manner as in a conventional transistor, resulting in a significant increase in the collector current.

The energy bands for the phototransistor are shown in Fig. 14.9. The photon-induced base current is returned to the collector through the emitter and the external circuitry. In so doing, electrons are supplied to the base region by the emitter where they are pulled into the collector by the electric field. The sensitivity of a phototransistor is a function of the collector–base diode quantum efficiency and also of the dc current gain of the transistor. Therefore, the overall sensitivity is a function of collector current.

When subjected to varying ambient temperature, collector current changes linearly with a positive slope of about 0.00667 per °C. The magnitude of this temperature coefficient is primarily a result of the increase in current gain versus temperature, since the collector–base photocurrent temperature coefficient is only about 0.001 per °C. The family of collector current versus collector voltage characteristics is very much similar to that of a conventional transistor. This implies that circuits with phototransistors can be designed by using the regular methods of

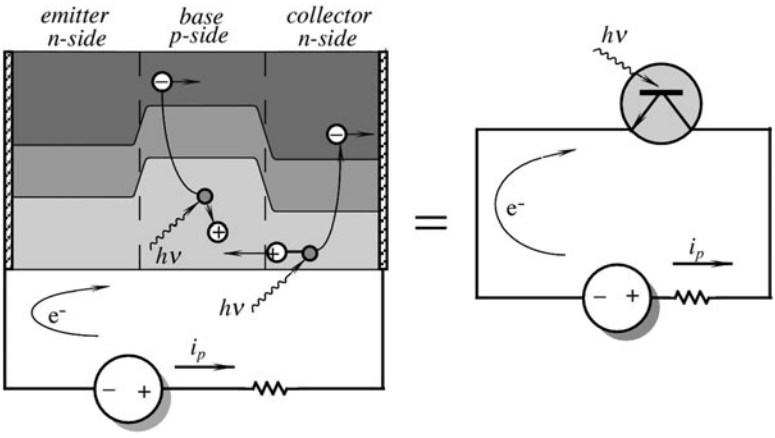

Fig. 14.9 Energy bands in a phototransistor

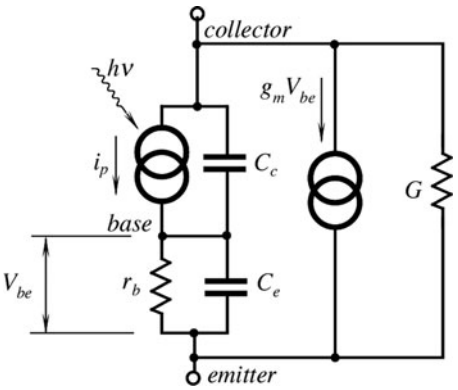

Fig. 14.10 Equivalent circuit of a phototransistor

transistor circuit techniques, except that its base should be used as an input of a photoinduced current that is supplied by its collector. Since the actual photogeneration of carriers occurs in the collector–base region, the larger the area of this region, the more carriers are generated; thus, the phototransistor is designed so to offer a large area to impinging light. A phototransistor can be either a two-lead or a three-lead device. In the latter case, the base lead is available, and the transistor may be used as a standard bipolar transistor with or without the additional capability of sensing light, thus giving a designer greater flexibility in circuit development. However, a two-lead device is the most popular as a dedicated photosensor.

When the base of the transistor is floating, it can be represented by an equivalent circuit shown in Fig. 14.10. Two capacitors C_c and C_e represent base–collector and base–emitter capacitances, which are the speed-limiting factors. Maximum frequency response of the phototransistor may be estimated from

$$f_1 \approx \frac{g_m}{2C_e}, \tag{14.13}$$

where f_1 is the current-gain-bandwidth product and g_m is the transistor's forward transconductance.

Whenever a higher sensitivity of a photodetector is required, especially if high-response speed is not of a concern, an integrated Darlington detector is recommended. It is comprised of a phototransistor whose emitter is coupled to the base of a bipolar transistor. Since a Darlington connection gives current gain equal to a product of current gains of two transistors, the circuit proves to be an efficient way to make a sensitive detector.

14.4 Photoresistors

As a photodiode, a photoresistor is a photoelectric device. It is a resistor whose resistance is called photoresistance R_p that changes as the function of incident light. The most common materials for its fabrication are cadmium sulfide (CdS) and

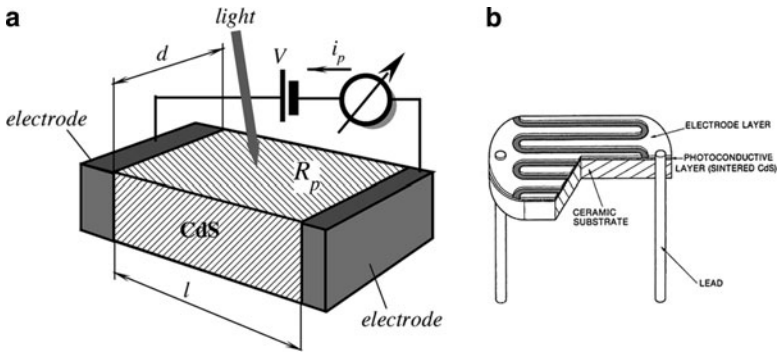

Fig. 14.11 Structure of a photoresistor (**a**) and a plastic-coated photoresistor having a serpentine shape (**b**)

cadmium selenide (CdSe), which are semiconductors whose resistances change upon light entering the surface. For its operation, a photoresistor requires a power source (excitation signal) because unlike a photodiode or phototransistor, it does not generate photocurrent – a photoeffect is manifested in change in the material's electrical resistance. Figure 14.11a shows a schematic diagram of a photoresistive cell. An electrode is set at each end of the photoconductor. In darkness, the resistance of the material is high. Hence, the applied voltage V results in small dark current, which is attributed to temperature effect. When light is incident on the surface, current i_p flows.

The reason for the current increase is the following. Directly beneath the conduction band of the crystal is a donor level and there is an acceptor level above the valence band. In darkness, the electrons and holes in each level are almost crammed in place in the crystal, resulting in a high resistance of the semiconductor.

When light illuminates the photoconductive crystal, photons are absorbed, which results in the added-up energy in the valence band electrons. This moves them into the conduction band, creating free holes in the valence band, increasing the conductivity of the material. Since near the valence band is a separate acceptor level that can capture free electrons not as easily as free holes, the recombination probability of the electrons and holes is reduced and the number of free electrons in the conduction band is high. Since CdS has a band gap of 2.41 eV, the absorption edge wavelength is $\lambda = c/\nu \approx 515$ nm, which is in the visible spectral range. Hence, the CdS detects light shorter than 515 nm wavelengths. Other photoconductors have different absorption edge wavelengths. For instance, CdS is the most sensitive at shorter wavelengths range, while Si and Ge are the most efficient in the near infrared.

The conductance of a semiconductor is given by

$$\Delta\sigma = ef(\mu_n\tau_n + \mu_p\tau_p). \tag{14.14}$$

where μ_n and μ_p are the free electron and hole movements (cm/V·s), τ_n and τ_p are the free electron and hole lives (sec), e is the charge of an electron, and f is the

number of generated carriers per second per unit of volume. For a CdS cell, $\mu_n\tau_n >> \mu_p\tau_p$; hence, conductance by free holes can be ignored. Then, the sensor becomes an n-type semiconductor. Thus,

$$\Delta\sigma = ef\mu_n\tau_n, \tag{14.15}$$

We can define sensitivity b of the photoresistor through a number of electrons generated by one photon (until the carrier lifespan ends):

$$b = \frac{\tau_n}{t_t}, \tag{14.16}$$

where $t_t = l^2/V\mu_n$ is the transit time for the electron between the sensor's electrodes, l is distance between the electrodes, and V is applied voltage. Then, we arrive at

$$b = \frac{\mu_n\tau_n V}{l^2}. \tag{14.17}$$

For example, if $\mu_n = 300\ \text{cm}^2/\text{V}\cdot\text{s}$, $\tau_n = 10^{-3}$ s, $l = 0.2$ mm, and $V = 1.2$ V, then the sensitivity is 900, which means that a single photon releases for conduction 900 electrons, making a photoresistor work as a photomultiplier. Indeed, a photoresistor is a very sensitive device.

It can be shown that for better sensitivity and lower cell resistance, a distance l between the electrodes should be reduced, while the width of the sensor d should be increased. This suggests that the sensor should be very short and very wide. For practical purposes, this is accomplished by fabricating a sensor in a serpentine shape (Fig. 14.11b).

Depending on the manufacturing process, the photoresistive cells can be divided into the sintered type, single crystal type, and evaporated type. Of these, the sintered type offers high sensitivity and easier fabrication of large sensitive areas, which eventually translated into lower cost devices. The fabrication of CdS cells consists of the following steps:

1. Highly pure CdS powder is mixed with appropriate impurities and a fusing agent.
2. The mixture is dissolved in water.
3. The solution in a form of paste is applied on the surface of a ceramic substrate and allowed to dry.
4. The ceramic subassemblies are sintered in a high-temperature oven to form a multicrystal structure. At this stage, a photoconductive layer is formed.
5. Electrode layers and leads (terminals) are attached.
6. The sensor is packaged into a plastic or metal housing with or without a window.

To tailor a spectral response of a photoresistor, the powder of step 1 can contain some variations; for instance, the addition of selenide or even the replacement of CdS for CdSe shifts the spectral response toward longer wavelengths (orange and red).

To illustrate how the photoresistors can be used, Fig. 14.12 shows two circuits. Circuit (A) shows an automatic light switch that turns lights on when illumination

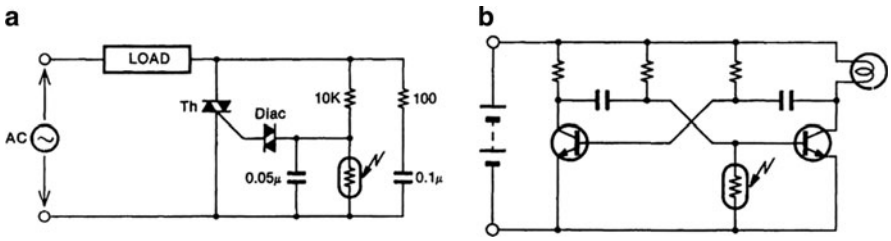

Fig. 14.12 Examples of photoresistor applications light switch (*left*) and beacon light (*right*) (courtesy of Hamamatsu Photonics K.K.)

drops (a turn-off part of the circuit is not shown). Circuit (B) shows a beacon with a free running multivibrator that is enabled at darkness when the resistance of a photoresistor becomes high. In these circuits, a photoresistor works as a resistor whose resistance is modulated by the light intensity.

14.5 Cooled Detectors

For measurements of objects emanating photons on the range of 2 eV or higher, quantum detectors having room temperature are generally used. For the smaller energies (longer wavelengths), narrower band gap semiconductors are required. However, even if a quantum detector has a sufficiently small energy band gap, at room temperatures its own intrinsic noise is much higher than a photoconductive signal. In other words, the detector will sense its own thermal radiation, and the useful signal will be buried in noise. Noise level is temperature dependent; therefore, when detecting long-wavelength photons, a signal-to-noise ratio may become so small that accurate measurement becomes impossible. This is the reason why for operation in the near- and far-infrared spectral ranges, a detector not only should have a sufficiently narrow energy gap, but its temperature has to be lowered to the level where intrinsic noise is reduced to an acceptable level.

Figure 14.13 shows typical spectral responses of some detectors with recommended operating temperatures. The operating principle of a cryogenically cooled detector is about the same as that of a photoresistor described in Sect. 14.4, except that it operates at far longer wavelengths at much lower temperatures. Thus, the sensor design becomes quite different. Depending on the required sensitivity and operating wavelength, the following crystals are typically used for this type of sensors: Lead sulfide (PbS), indium arsenide (InAs), germanium (G), lead selenide (PbSe), and mercury–cadmium–telluride (HgCdTe).

Cooling shifts responses to longer wavelengths and increases sensitivity. However, response speeds of PbS and PbSe become slower with cooling. Methods of cooling include Dewar cooling using dry ice, liquid nitrogen, liquid helium (Fig. 14.14), or thermoelectric coolers operating on the Peltier effect (see Sect. 3.9.2).

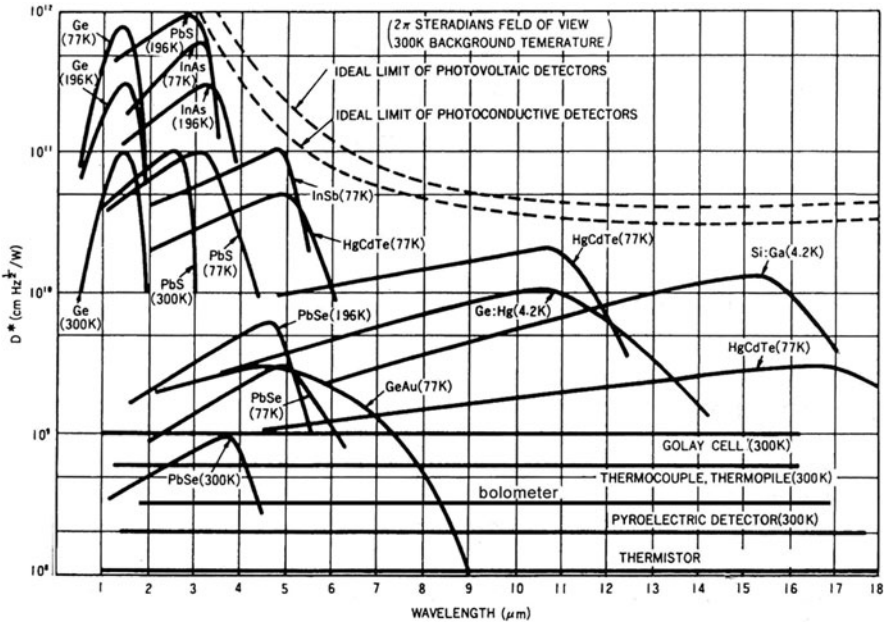

Fig. 14.13 Operating ranges for some infrared detectors

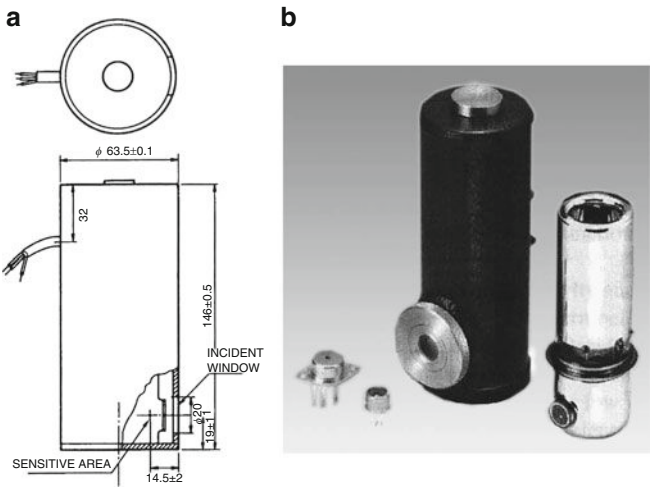

Fig. 14.14 Cryogenically cooled MCT quantum infrared detectors dimensional drawing of a dewar type (in mm) (**a**) and outside appearances of canned and dewar detectors (**b**) (courtesy of Hamamatsu Photonics K.K.)

Table 14.2 Typical specifications for MCT-far-infrared detectors

Sensitive area (mm)	Temperature (°C)	l_p (μm)	l_c (μm)	FOV (°)	Dark resist (kΩ)	Rise time (μs)	Max current (mA)	D* @ l_p
1 × 1	−30	3.6	3.7	60	1	10	3	10^9
1 × 1	−196	15	16	60	20	1	40	3×10^9

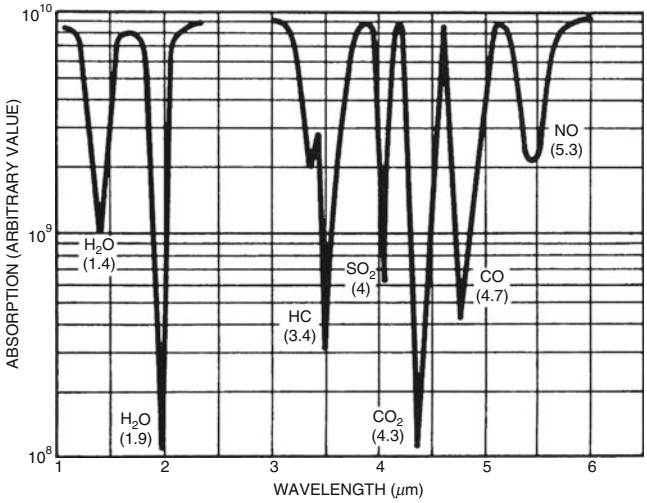

Fig. 14.15 Absorption spectra of gaseous molecules

As an example, Table 14.2 lists typical specifications for an MCT photoconductive detector. MCT stands for the mercury–cadmium–telluride type of a sensing element.

Applications of the cryogenically cooled quantum detectors include measurements of optical power over a broad spectral range, thermal temperature measurement and thermal imaging, detection of water content, and gas analysis.

Figure 14.15 depicts the gas absorption spectra for various molecules. Water strongly absorbs at 1.1, 1.4, 1.9, and 2.7 μm. Thus, if it is required to determine the moisture content, for example, in coal, the monochromatic light is projected on both the test and reference samples. The reflected light is detected, and the ratio is calculated for the absorption bands. The gas analyzer makes use of absorption in the infrared region of the spectrum. This allows fabrication of the spectrophotometers that measure the gas density. Also, similar sensors can be employed for measuring the automobile exhaust gases (CO, HC, CO_2), for emission control (CO, SO, NO_2), detecting fuel leakage (CH_4, C_3H_2), etc.

14.6 Image Sensors

The charge-coupled device (CCD)[1] and complementary metal oxide semiconductor (CMOS) image sensors are two different technologies presently used for capturing images digitally. Each has unique strengths and weaknesses giving advantages in different applications.

Both types of imagers convert light into electric charge and process it into electronic signals. In a CCD sensor, every pixel's charge is transferred through a very limited number of output nodes (often just one) to be converted to voltage, buffered, and sent off-chip as an analog signal. Then, the analog signal is digitized by an external A/D converter. All of the pixels can be devoted to light capture, and the output's uniformity (a key factor in image quality) is high. In a CMOS sensor, each pixel has its own charge-to-voltage conversion, and the sensor often also includes amplifiers, noise correction, and digitization circuits, so that the chip outputs digital bits. These other functions increase the design complexity and reduce the area available for light capture. With each pixel doing its own conversion, uniformity is lower. But the chip can be built to require less off-chip circuitry for basic operation.

CMOS imagers offer more integration (more functions on the chip), lower power dissipation (at the chip level), and the possibility of smaller system size, but they have often required trade-offs between image quality and device cost. CMOS cameras may require fewer components and less power, but they still generally require companion chips to optimize image quality, increasing cost and reducing the advantage they gain from lower power consumption. CCD devices are less complex than CMOS, so they cost less to design. CCD fabrication processes also tend to be more matured and optimized; in general, it will cost less (in both design and fabrication) to yield a CCD than a CMOS imager for a specific high-performance application. The choice continues to depend on the application and the vendor more than the technology.

CCDs and CMOS imagers were both invented in the late 1960s and 1970. CCD became dominant, primarily because they gave far superior images with the fabrication technology available. CMOS image sensors required more uniformity and smaller features that the silicon wafer foundries could not produce at the time. Nowadays, renewed interest in CMOS was based on expectations of lowered power consumption, camera-on-a-chip integration, and lowered fabrication costs.

Both CCDs and CMOS imagers can offer excellent imaging performance when designed properly. CCDs have traditionally provided the performance benchmarks in the photographic, scientific, and industrial applications that demand the highest image quality (as measured in quantum efficiency and noise) at the expense of system size. Astronomers say that CCDs have a high-quantum efficiency (QE), meaning that a large percentage of incoming photons are actually detected. While photographic plates might capture *one* photon out of every 100, modern CCDs

[1] In 2009, Willard S. Boyle and George E. Smith received a Nobel Prize for their invention of CCD in 1969.

would capture 80 photons out of every 100. This allows for a substantial decrease in exposure time. CCDs are also linear in nature, meaning that the signal they produce is directly proportional to the amount of light collected. This makes it easier to calculate the number of photons that hit the detector in the time of an exposure.

14.6.1 CCD Sensor

A CCD chip is divided into pixels. Each pixel has a potential well that collects the electrons produced by the photoelectric effect. At the end of an exposure (frame), each pixel has collected an amount of electrons (i.e., charge) proportional to the amount of light that fell onto it. The CCD is then read out by cycling the voltages applied to the chip in a process called "clocking." Due to the structure of a CCD, clocking causes the charge in one pixel to be transferred to an adjacent pixel. To understand how the whole chip can operate, consider the following analogy suggested by Jerome Kristian and shown in Fig. 14.16.

The incoming photons are represented by the raindrops, and the CCD chip is a 2D array of buckets. Each bucket represents a pixel, and the water it collects is the combined charge accumulation due to photoelectrons. Once the rain has stopped (the shutter is closed), conveyor belts move the columns of buckets down one row (the gates are clocked). The water in the buckets at the edge of the array pours into more buckets on a horizontal conveyor belt. This conveyor belt then pours these buckets one at a time into a container on a scale, which is a graduated cylinder (the readout electronics). The volume of water from each bucket is measured and rounded to the nearest milliliter [corresponding to the digital output of a CCD, which reports the counts, or analog-to-digital units (ADUs), from each pixel]. Then, the image is formed by reconstructing the distribution of rainfall on the array.

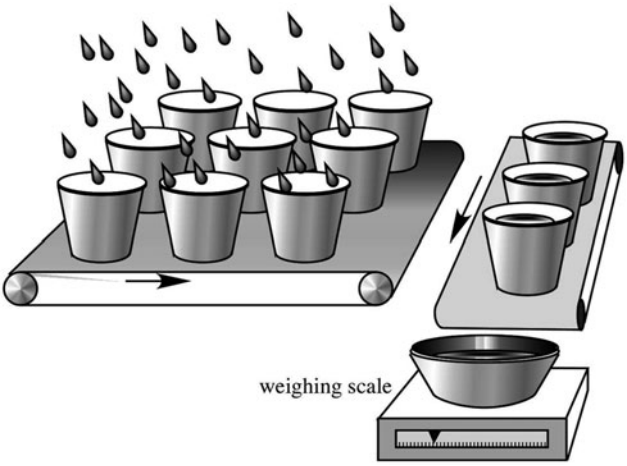

Fig. 14.16 Analogy of the CCD operation

Most CCDs have a setting called binning that causes them to read out in a slightly different manner. When binning is used, blocks of pixels are grouped together into "superpixels." Each superpixel acts as one large pixel, with several results. Read times are faster, since fewer actual measurements of charge are performed. Each superpixel is also more sensitive, since it can collect more photons for a given exposure time, but this comes at the cost of resolution.

14.6.2 CMOS-Imaging Sensors

Like CCDs, these imagers are made from silicon, but as the name implies, the process they are made in is called CMOS. This process is today the most common method of making processors and memories, meaning CMOS imagers take advantage of the process and cost advancements created by these other high-volume devices. Because CMOS imagers are created in the same process as processors, memories, and other major components, CMOS imagers can integrate with these same components onto a single piece of silicon. Like CCDs, CMOS imagers include an array of photosensitive diodes (PD), one diode within each pixel. Unlike CCDs, however, each pixel in a CMOS imager has its own individual amplifier integrated inside (Fig. 14.17). Since each pixel has its own amplifier, the pixel is referred to as an "active pixel." In addition, each pixel in a CMOS imager can be read directly on an *x*–*y* coordinate system, rather than through the "bucket-brigade" process of a CCD. This means that while a CCD pixel always transfers a charge, a

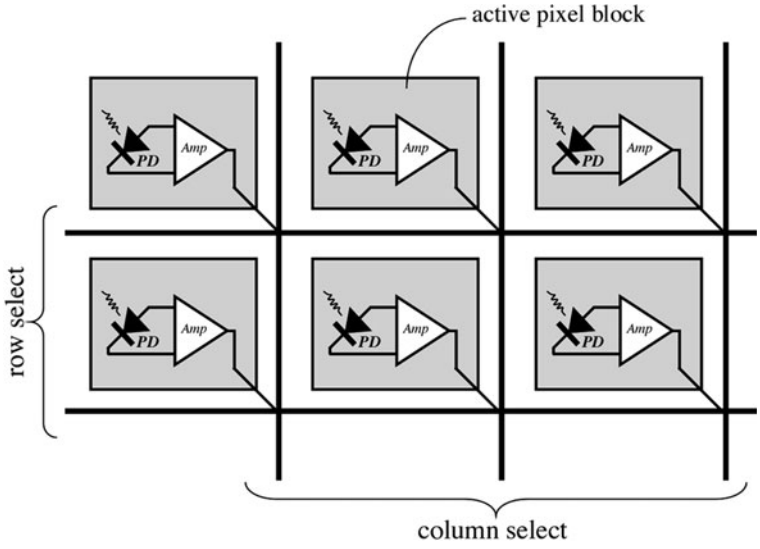

Fig. 14.17 Organization of a CMOS-imaging sensor

CMOS pixel always detects a photon directly, converts it to a voltage, and transfers the information directly to the output.

14.7 Thermal Detectors

Thermal-infrared detectors are primarily used for detecting infrared radiation in mid- and far-infrared spectral ranges and noncontact temperature measurements, which have been known in industry for about 60 years under the name of *pyrometry* from the Greek word $\pi\upsilon\rho$ (fire). The respective thermometers are called *radiation* pyrometers. Today, noncontact methods of temperature measurement embrace a very broad range, including subzero temperatures, which are quite far away from that of flame. Therefore, it appears that *radiation thermometry* is a more appropriate term for this technology.

A typical infrared temperature sensor consists of the following:

1. A *sensing element* – a component that is responsive to electromagnetic radiation in the selected wavelength range from visible to far infrared, depending on the temperature range of the instrument. The main requirements to the element are fast, predictable and strong response to thermal radiation, and a good long-term stability.
2. A *supporting structure* to hold the sensing element and to expose it to thermal radiation. The structure should have low thermal conductivity to minimize heat loss.
3. A *housing* that protects the sensing element from the environment. It usually should be hermetically sealed and often filled with either dry air or inert gas such as argon or nitrogen.
4. A *protective window* that is impermeable to environmental factors and transparent in the wavelength of detection. The window may have surface coatings to improve transparency and to filter out undesirable portions of the spectrum.

Below the mid-infrared range, thermal detectors are much less sensitive than quantum detectors. Thus, the thermal IR sensors are used for the wavelengths longer than 1 μm. Their operating principle is based on a sequential conversion of thermal radiation into heat, and then, conversion of heat magnitude or heat flow into an electrical signal by employing conventional methods of heat detection. In other words, a thermal-radiation sensor in its operating principle resembles a conventional temperature sensor (see Chap. 16) that converts temperature into electric signal. The higher the temperature increase, the higher the electrical output. This sets the design criterion – make the IR-sensing element to respond with high-temperature increase or decrease as practical and then measure that increase with a conventional, albeit specially designed, temperature sensor like a thermocouple or thermistor.

According to (3.133), the infrared flux that is absorbed by a thermal detector is proportional to a geometry factor A, which for a uniform spatial distribution of

radiation is equal to the sensor's area. For instance, if a thermal-radiation-sensing element of a 5 mm^2 surface area and ideal absorptivity at the initial temperature 25°C is placed inside an enclosed chamber whose temperature is 100°C, the sensor will receive the initial radiative power of 3.25 mW. Depending on the element's thermal capacity, its temperature will rise exponentially until thermal equilibrium between the element and its environment occurs. That is, the element will reach temperature of 100°C because its thermal coupling to the enclosed chamber is total – 100%. In practice, the sensing element's temperature never reaches that of an object because a thermal coupling between the sensing element and the object is never 100%. Unlike a hypothetical sensing element that is placed inside a radiative chamber, a real-sensing element is rather poorly thermally coupled to the heat source. The sensing element is usually positioned inside the radiation thermometer that may be at ambient temperature, while the object is remote and may be either very hot or very cold.

While the sensing element exchanges heat by radiation, a substantial portion of the absorbed heat is lost through a supporting structure and wires, as well as through gravitational convection and also through stray radiation. Thus, the equilibrium temperature is always somewhere in-between the object's temperature and the initial temperature of the thermal-sensing element.

All thermal-radiation detectors can be divided into two classes: passive-infrared (PIR) and active-far-infrared (AFIR) detectors. Passive detectors absorb incoming radiation and convert it to heat, while active detectors generate heat from the excitation circuit and measures heat loss that depends on the environment and the object.

14.7.1 Golay Cells

The Golay cells are the broadband detectors of infrared radiation. They are extremely sensitive, but also extremely delicate. The operating principle of the cell is based on the detection of a thermal expansion of gas trapped inside an enclosure. This is why these detectors sometimes are called thermopneumatic detectors. Figure 14.18 depicts an enclosed chamber having two membranes – the upper and lower. The upper membrane is coated with a radiative heat absorber (e.g., goldblack), while the lower membrane has a mirror surface (e.g., coated with aluminum).

The mirror is illuminated by a light source. The incident light beam is reflected from the mirror and impinges on a position-sensitive detector (PSD). The upper membrane is exposed to infrared radiation that is absorbed by the coating and elevates temperature of the membrane. This, in turn, warms up gas that is trapped inside the sensor's chamber. Gas expands and its pressure goes up. The increase in the internal pressure deflects the lower membrane that bulges out. A change in the mirror curvature deflects the reflected light beam. The reflected light impinges on the PSD at various locations, depending on the degree of bulging of the membrane, and therefore of the intensity of the absorbed radiation. The entire sensor may be micromachined using modern MEMS technology (see Chap. 18). The degree of the

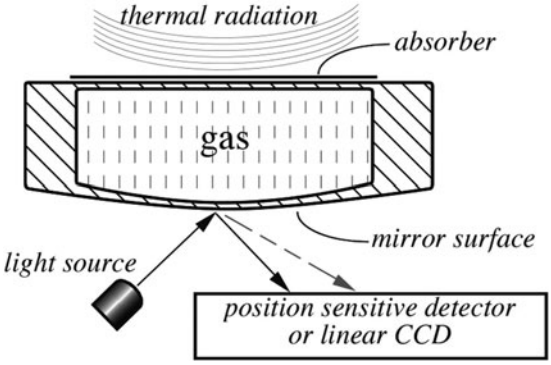

Fig. 14.18 Golay cell detector for mid- and far-infrared radiation

lower membrane deflection alternatively may be measured by different methods, for example, by using a FP interferometer (see Sect. 7.4.4).

14.7.2 Thermopile Sensors

Thermopiles belong to a class of PIR detectors. Their operating principle is the same as that of a thermocouple. A single thermocouple is a low-sensitivity device responding with tens of microvolts per 1°C change. In a thermal-radiation sensor, temperature change of the sensing element when exposed to the object is very small – as low as 0.001°C. Thus, a stronger sensor response is required. This is achieved by increasing a number of thermocouples that makes a thermopile (piling up). A thermopile is a chain of serially connected thermocouples, typically 50–100 junctions. The chain will produce a 50–100 times stronger signal when properly connected and used. Originally, it was invented by Joule to increase the output signal of a thermoelectric sensor. He connected several thermocouples in a series and thermally joined together their "hot" junctions. Nowadays, thermopiles have a different configuration. Their prime application is thermal detection of light in the mid- and far-infrared spectral ranges.

An equivalent schematic of a thermopile sensor is shown in Fig. 14.19a. The sensor consists of a frame having a relatively large thermal mass, which is the place where the "cold" junctions are positioned. The frame may be thermally coupled with a reference temperature sensor or attached to a thermostat having a precisely known temperature. The frame supports a thin membrane whose thermal capacity is very small due to its geometry. A very small thermal capacity results in a larger temperature increase when subjected to thermal radiation [see (3.133) in Chap. 3]. The membrane surface carries the "hot" junctions of thermocouples. Words "hot" and "cold" are the remnants of the traditional thermocouple jargon and used here conditionally since the junctions in reality are rarely cold or hot.

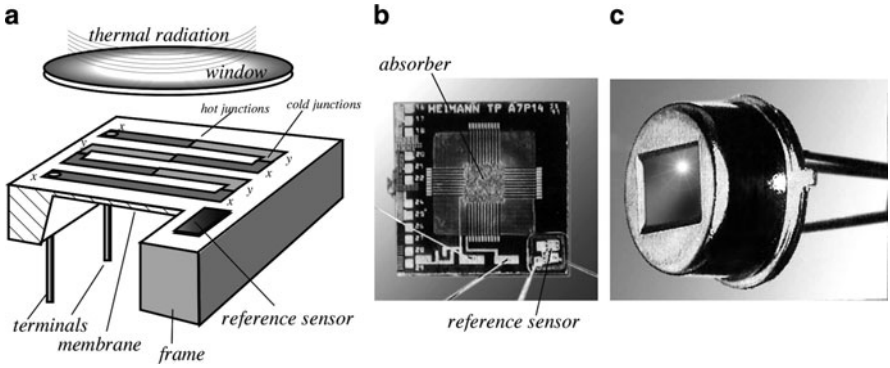

Fig. 14.19 Thermopile for detecting thermal radiation, equivalent schematic with a reference temperature sensor attached, *x* and *y* are different materials (**a**); micromachined thermopile sensor. Note the semiconductor reference temperature sensor on the silicon frame where the cold junctions are deposited, and the absorptive coating on the hot junctions in the center of the membrane (**b**); and sensor in a TO-5 packaging (**c**)

Infrared light is absorbed by or emanated from the membrane, and, in response, temperature of the membrane changes. Since the membrane carries "hot" junctions, the temperature differential with respect to the "cold" junctions located on the frame generate thermoelectric voltage. The membrane temperature increase depends on the thermal capacity, thermal conductivity to the frame, and intensity of the infrared light.

The best performance of a thermopile is characterized by high sensitivity and low noise, which may be achieved by the junction materials having high-thermoelectric coefficient α, low-thermal conductivity, and low-volume electric resistivity. Besides, the "hot" and "cold" junction pairs should have thermoelectric coefficients of the opposite signs. This dictates the selection of the materials. Unfortunately, most of metals having low-electrical resistivity (gold, copper, and silver) have only very poor thermoelectric coefficients. The higher electrical resistivity metals (especially bismuth and antimony) possess high-thermoelectric coefficients, and they are often selected for designing thermopiles. By doping these materials with Se and Te, the thermoelectric coefficient was improved up to 230 $\mu V\ K^{-1}$ [5], and the original thermopiles were constructed with these metals.

Methods of construction of the metal junction thermopiles may differ to some extent, but all incorporate vacuum deposition techniques and evaporation masks to apply the thermoelectric materials, such as bismuth and antimony. The number of junctions varies from 20 to several hundreds for special designs. The "hot" junctions are often coated with absorber of thermal radiation. For example, they may be blackened (with Nichrome,[2] goldblack, or organic paint) to improve their absorptivity of the infrared radiation.

[2]Alloy of 80% nickel and 20% chromium has emissivity (absorptivity) over 0.80.

Table 14.3 Typical specifications of a thermopile

Parameter	Value	Unit	Conditions
Sensitive area	0.5–2	mm^2	–
Responsivity	50	V/W	6–14 μm, 500 k
Noise	30	$nV/\sqrt{Hz}$	25°C, rms
Equivalent resistance	50	kΩ	25°C
Thermal time constant	60	ms	–
TCR	0.15	%/K	–
Temperature coefficient of responsivity	−0.2	%/K	–
Operating temperature	−20 to +80	°C	–
Storage temperature	−40 to 100	°C	–
Price	1–10	US$	–

A thermopile is a dc device whose output voltage follows its "hot" junction temperature almost linearly. A thermopile can be modeled as a thermal flux-controlled voltage source that is connected in series with a fixed resistor. The sensor is hermetically sealed in a metal can with a hard infrared transparent window, for instance, silicon, germanium, or zinc selenide (Fig. 14.19c). The output voltage V_s is nearly proportional to the incident radiation. The thermopile-operating frequency limit is mainly determined by the thermal capacity and thermal conductivity of the membrane, which are manifested through a thermal time constant. The sensor exhibits quite a low noise that is equal to the thermal noise of the sensor's equivalent resistance, which is of 20–100 kΩ. Typical properties of a metal-type thermopile sensor are given in Table 14.3.

The output signal of a thermopile sensor depends on a temperature gradient between the source of the thermal radiation and the sensing surface. As a result, the transfer function of a thermopile is a three-dimensional surface whose shape is governed by the Stefan–Boltzmann law (see Fig. 2.3).

Nowadays, bismuth and antimony are being replaced by the silicon thermopiles. These thermopiles are more efficient and reliable [6]. Table A.11 in Appendix lists thermoelectric coefficients for selected elements. It is seen that the coefficients for crystalline and polycrystalline silicon are very large, and the volume resistivity is relatively low. The advantage of using silicon is in the possibility of employing standard IC processes that results in a significant cost reduction. The resistivity and the thermoelectric coefficients can be adjusted by the doping concentration. However, the resistivity increases much faster, and the doping concentration has to be carefully optimized for the high-sensitivity-low-noise ratios.

Figure 14.19b shows a semiconductor thermopile sensor that was fabricated by employing a micromachining (MEMS) technology. The central part of the silicon substrate is removed by means of anisotropic etching from the back, leaving only about 1 μm thin sandwich layer (membrane) of $SiO_2 - Si_3N_4$ on top, which has low-thermal conductivity. Onto this membrane, thin conductors of two different thermoelectric materials (polysilicon and aluminum) are deposited. This allowed the production of sensors with the negligible temperature coefficient of sensitivity, which is an important factor for operation over broad ambient temperatures ranges.

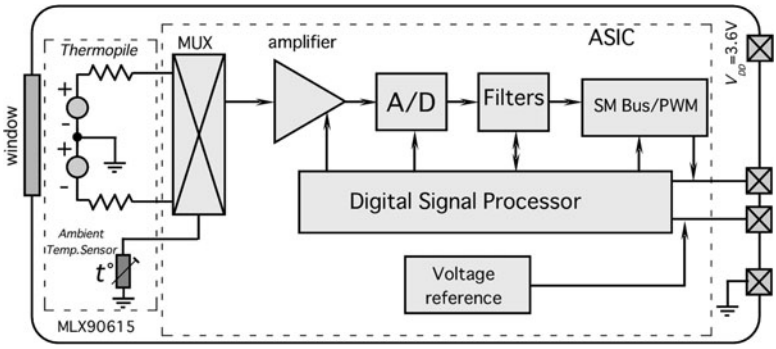

Fig. 14.20 Block-diagram of an integrated IR thermometer with a thermopile

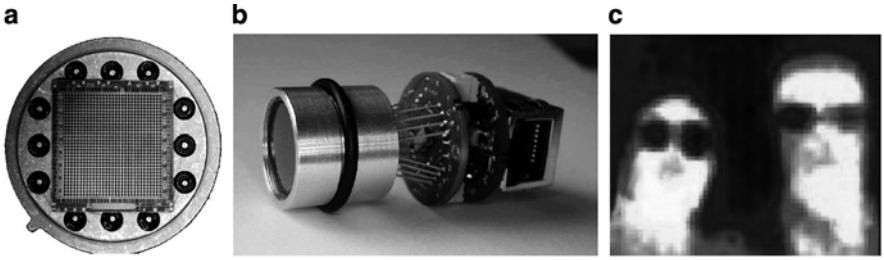

Fig. 14.21 Thermopile thermal-imaging sensor: sensing surface (**a**); imaging module (**b**); and example of a thermal image (**c**) (courtesy of Heimann Sensors, www.heimannsensor.com)

The modern trend in the IR-sensing technology is integrating a thermopiles sensor together with an amplifier, A/D converter, and other processing circuitry. A Belgium company Melexis (www.melexis.com) developed an entire IR thermometer MLX90615 in a miniature TO-46 can that contains the thermopile and data processing ASIC chips (Fig. 14.20). The ambient sensor is required for computing the surface temperature of an object. A small output signal from the thermopile is fed into a precision amplifier having an offset voltage as small as 0.5 μV. The digital signal processor (DSP) outputs the measured temperature or individual outputs from the sensor.

It can be said that the above described thermopile is a single-pixel thermal-radiation sensor. Yet, a sensor with multiple thermopile pixels can be designed and used for a simultaneous detection of thermal radiation from multiple sources or for thermal imaging. An example of such a sensor is shown in Fig. 14.21 where the thermopile pixels are arranged in a 32 × 31 matrix. The number of junctions per pixel is 80, and the thermoelectric junction material is n-poly/p-poly Si. The pixels have a size of 150 μm and are positioned at a distance of 220 μm from one another. The sensing module HTPA32x31 from Heimann Sensors has the imbedded pre-amplifiers, multiplexer, and A/D converter. Advantage of the module is that it does

not require a cryogenic cooling and operates over a broad ambient temperature range.

14.7.3 Pyroelectric Sensors

A pyroelectric sensor belongs to a class of PIR detectors. Unlike a thermopile, this sensor can be called an ac sensor as it is only responsive to a change in thermal-radiation signal and not responsive to the signal magnitude. Refer to Sect. 3.7 for description of the pyroelectric effect. The pyroelectric element consists of three essential components: The pyroelectric ceramic plate and two electrodes deposited on the opposite sides of the plate. The plate is supported inside the sensor's housing with as little contact area with the housing as possible. A typical construction of a pyroelectric sensor is shown in Fig. 14.22a. It is housed in a metal TO-5 or TO-39 can for better shielding and is protected from the environment by a silicon or any other appropriate window that is substantially transparent in the mid- and far-infrared spectral range. The inner space of the can is filled with dry air or nitrogen. Usually, two identical sensing elements are oppositely, serially, or in parallel connected for the better compensation of rapid thermal changes and mechanical stresses resulting from acoustical noise and vibrations. One of the elect is coated with either heat-absorbing paint or goldblack or is fabricated of nichrome, while the electrode on the other element is either shielded from radiation or gold-plated for better reflectivity that prevents absorption of the IR signal. Nichrome has high emissivity (absorptivity) and thus serves a dual purpose – collects electric charge from the plate and absorbs thermal radiation. For applications in the PIR motion detectors (see Sect. 6.5.3), both pyroelectric elements are exposed to the window and are nearly identical.

A dual element is often fabricated on a single plate (flake) of a crystalline material (Fig. 14.22b). The metallized pattern on both sides of the flake forms two serially connected capacitors C_1 and C_2. Figure 14.22c shows an equivalent circuit of a dual pyroelectric element. This design has the benefit of a good balance of both elements, resulting in a better rejection of common-mode interferences.

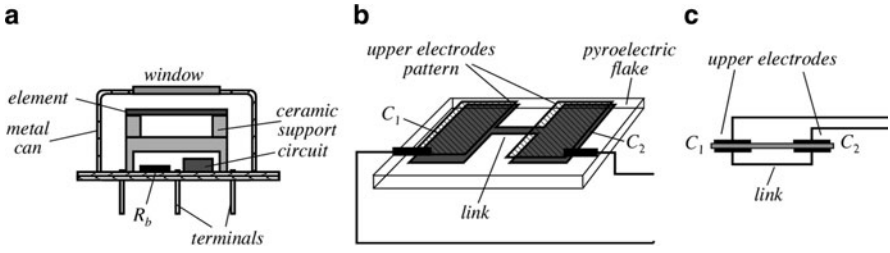

Fig. 14.22 A dual pyroelectric sensor assembled in a metal can (**a**); metal electrodes are deposited on the opposite sides of a pyroelectric plate (**b**); and equivalent circuit of a dual element (**c**)

Note that the sensing element exists only between the opposite electrodes, and the portion of the flake that is not covered by the electrodes does not participate in generation of a useful signal. A major problem in designing a pyroelectric detector is in its sensitivity to mechanical stress and vibrations. All pyroelectrics are also piezoelectrics; therefore, while being sensitive to thermal radiation, the pyroelectric sensors are susceptible to interferences that sometimes are called "microphonics." For better noise rejection, the crystalline element must be mechanically decoupled from the outside, especially from the terminals and the metal can.

A pyroelectric element has a very high internal resistance. High but not infinitely high, so that resistance forms a leakage resistor connected in parallel with the capacitor C_1 and likewise with C_2. The value of that resistor is on the orders of 10^{12}–10^{14} Ω. In a practical application, the sensor is connected to the circuit that contains a bias resistor R_b and an impedance converter ("circuit" in Fig. 14.22a). The converter may be either a voltage follower (for instance, a JFET transistor) or a current-to-voltage converter. The voltage follower (Fig. 14.23a) converts a high-output impedance of the sensor (combined capacitance C in parallel with a bias resistance R_b) into the output resistance of the follower, which in this example is determined by the transistor's transconductance in parallel with 47 kΩ. The advantage of this circuit is in its simplicity, low cost, and low noise. A single JFET follower is the most cost effective and simple; however, it suffers from two major drawbacks. The first is the dependence of its speed response on the so-called electrical time constant, which is a product of the sensor's combined capacitance C and the bias resistor R_b

$$\tau_e = CR_b. \tag{14.18}$$

For example, a typical dual sensor may have $C = 40$ pF and $R_b = 50$ GΩ, which yield $\tau_e = 2$ s, corresponding to a first-order frequency response with the upper cutoff frequency at 3 dB level equal to about 0.08 Hz, a very low frequency indeed. This makes the voltage follower suitable only for limited applications, where speed response in not too important. An example is the detection of movement of people (see Chap. 6). The second drawback of the circuit is a large offset voltage across the

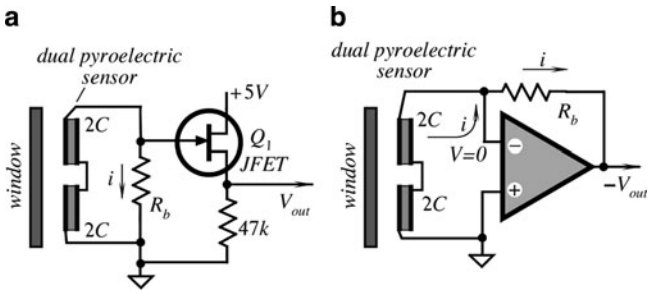

Fig. 14.23 Impedance converters for pyroelectric sensors voltage follower with JFET (**a**) and current-to-voltage converter with operational amplifier (**b**)

output resistor. This voltage depends on the type of the transistor and is temperature dependent. Thus, the output V_{out} is the sum of two voltages: The offset voltage that can be as large as several volts, and the alternate pyroelectric voltage that may be on the order of millivolts.

A more efficient, however, more expensive circuit for a pyroelectric sensor is an *I*/*V* (current-to-voltage) converter (Fig. 14.23b). Its advantage is in a faster response and insensitivity to the capacitance of the sensor element. The sensor is connected to an inverting input of the operational amplifier, that possesses properties of the so-called virtual ground (the similar circuits are shown in Figs. 14.6 and 14.8). That is, the voltage in the inverting input is constant and almost equal to that of a non-inverting input, which in this circuit is grounded. Thus, the voltage across the sensor is forced by the feedback to stay near zero; so, the combined capacitor *C* has no chance to charge. The output voltage follows the shape of the electric current (a flow of charges) generated by the sensor (Fig. 3.29). The circuit should employ an operational amplifier with a very low-bias current (on the order of 1 pA). There are three major advantages in using this circuit: A fast response, insensitivity to the capacitance of the sensor, and a low-output offset voltage. However, being a broad bandwidth circuit, a current-to-voltage converter may suffer from higher noise.

At very low frequencies, both circuits, the JFET and *I*/*V* converter, transform pyroelectric current i_p into output voltage. According to Ohm's law

$$V_{out} = i_p R_b. \tag{14.19}$$

For instance, for the pyroelectric current of 10 pA (10^{-11} A) and the bias resistor of 5×10^{10} Ω (50 Gigohm), the output voltage swing is 500 mV. Either the JFET transistor or the operational amplifier must have low-input bias currents (I_B) over an entire operating temperature range. The CMOS OPAMs are generally preferable as their bias currents are on the order of 1 pA.

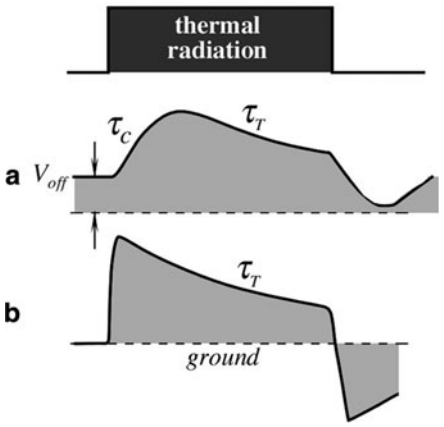

Fig. 14.24 Output signals of a voltage follower (**a**) and current-to-voltage converter (**b**) in response to a step function of thermal radiation

It should be noted that the circuits described above (Fig. 14.23) produce output signals of quite different shapes. The voltage follower's output voltage is a repetition of voltage across the element and R_b (Fig. 14.24a). It is characterized by two slopes: The leading slope having an electrical time constant $\tau_e = CR_b$, and the decaying slope having thermal time constant τ_T. Voltage across the element in the current-to-voltage converter is essentially zero and, contrary to the follower, the input impedance of the converter is very low. In other words, while the voltage follower acts as a voltmeter, the current-to-voltage converter acts as an amperemeter. The leading edge of its output voltage is fast (determined by a stray capacitance across R_b), and the decaying slope is characterized by τ_T. Thus, the converter's output voltage repeats the shape of the sensor's pyroelectric current (Fig. 14.24b).

Fabrication of Gigohm-range resistors that are essential for use with pyroelectric sensors is not a trivial task. High-quality bias resistors must have good environmental stability, low-temperature coefficient of resistance (TCR), and low-voltage coefficient of resistance (VCR). The VCR is defined as

$$\xi = \frac{R_1 - R_{0.1}}{R_{0.1}} \cdot 100\%, \tag{14.20}$$

where R_1 and $R_{0.1}$ are the resistances measured, respectively, at 1 and 0.1 V. Usually, VCR is negative, that is, the resistance value drops with an increase in voltage across the resistor (Fig. 14.25a). Since the pyroelectric sensor's output is proportional to the product of the pyroelectric current and the bias resistor, VCR results in nonlinearity of an overall transfer function of the sensor plus the circuit. A high-impedance resistor is fabricated by depositing a thin layer of a semiconductive ink on a ceramic (alumina) substrate, firing it in a furnace, and subsequent covering of the surface with a protective coating. A high quality, relatively thick (at least 50 μm thick) hydrophobic coating is very important for protection against moisture, since even a small amount of water molecules may cause oxidation of the semiconductive layer. This causes a substantial increase in the resistance and a poor long-term stability. A typical design of a high-impedance resistor is shown in Fig. 14.25b.

In applications where high accuracy is not required, such as thermal-motion detection, the bias resistor can be replaced with one or two zero-biased parallel-opposite connected silicon diodes.

For the detection of a thermal radiation, a distinction exists between two cases in which completely different demands have to be met with respect to a pyroelectric material and its thermal coupling to the environment [7]:

1. Fast sensors detect radiation of high intensity but very short duration (nanoseconds) of laser pulses, with a high repetition on the order of 1 MHz. The sensors are usually fabricated from single-crystal pyroelectrics such as lithium tantalate ($LiTaO_3$) or triglycine sulfate (TGS). This assures a high linearity of response. Usually, the materials are bonded to a heat sink.

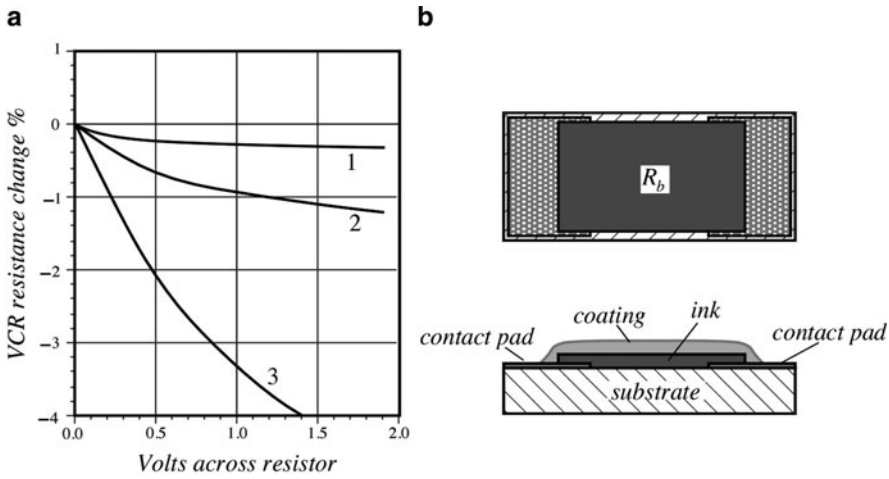

Fig. 14.25 High-impedance resistor VCRs for three different types of the resistor (**a**) and semiconductive ink is deposited in the alumina substrate (**b**)

2. Sensitive sensors detect thermal radiation of low intensity, however, with a relatively low rate of change. Examples are infrared thermometry and motion detection [8–10]. These sensors are characterized by a sharp temperature rise in the field of radiation. This generally requires a good thermal coupling with a heat source. Optical devices such as focusing lenses and waveguides are generally employed. A heat transfer to the environment (sensor's housing) must be minimized. If well designed, such a sensor can have a sensitivity approaching that of a cryogenically cooled quantum detector [7]. Commercial pyroelectric sensors are implemented on the basis of single crystals such as TGS and LiTaO3, or lead zirconate titanate (PZT) ceramics. PVDF film is also occasionally used thanks to its high-speed response and good lateral resolution.

14.7.4 Bolometers

Bolometers are miniature RTDs or thermistors (see Sect. 16.3), or other temperature-sensitive resistors that are mainly used for measuring rms values of electromagnetic radiation over a very broad spectral range from mid infrared to microwaves. Applications include infrared temperature detection and imaging, measurements of local fields of high power, the testing of microwave devices, RF antenna beam profiling, testing of high power microwave weapons, monitoring of medical microwave heating, and others. The operating principle is based on a fundamental relationship between the absorbed electromagnetic signal and dissipated power [11]. The conversion steps in a bolometer are as follows:

1. An ohmic resistor is exposed to electromagnetic radiation. The radiation is absorbed by the resistor and converted into heat.
2. The heat elevates resistor's temperature above the ambient.
3. The temperature increase reduces the bolometer's ohmic resistance.

A temperature increase is a representation of the electromagnetic power. Naturally, this temperature differential can be measured by any suitable method. These methods are covered in Chap. 16. Here, we just briefly outline the most common methods of bolometer fabrications, which evolved quite dramatically since Langley first invented a bolometer over 100 years ago.

A basic circuit for the voltage-biased-bolometer application is shown in Fig. 14.26a. It consists of a bolometer (a temperature-sensitive resistor) having resistance R, a stable reference resistor R_0, and a bias voltage source E. The voltage V across R_0 is the output signal of the circuit. It has the highest value when both resistors are equal. Sensitivity of the bolometer to the incoming electromagnetic (EM) radiation can be defined as [12]

$$\beta_V = \frac{\alpha \varepsilon Z_T E}{4\sqrt{1 + (\omega\tau)^2}}, \tag{14.21}$$

where $\alpha = (dR/dT)/R$ is the temperature coefficient of resistance (TCR) of the bolometer, ε is the surface emissivity, Z_T is the bolometer thermal resistance, which depends on its design and the supporting structure, τ is the thermal time constant, which depends on Z_T and the bolometer's thermal capacity, and ω is the frequency.

Since bolometer's temperature increase, ΔT is

$$\Delta T = T - T_0 \approx P_E Z_T = \frac{E^2}{4R} Z_T, \tag{14.22}$$

and the resistance of RTD bolometer can be represented by a simplified (16.14)

$$R = R_0(1 + \alpha_0 \Delta T), \tag{14.23}$$

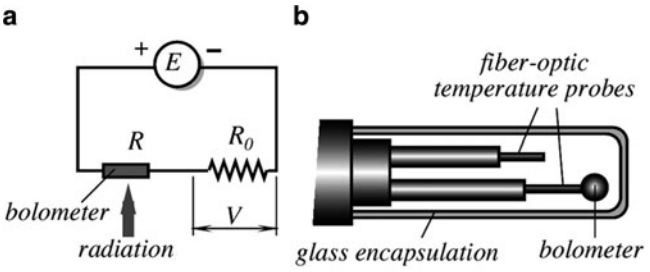

Fig. 14.26 Equivalent circuit of electrically biased bolometer (**a**) and a design of an optical bolometer (**b**)

then (14.21) can be rewritten as

$$\beta_V = \frac{1}{2}\varepsilon\alpha\sqrt{\frac{R_0 Z_T \Delta T}{(1+\alpha_0\Delta T)\left[1+(\omega\tau)^2\right]}} \tag{14.24}$$

Therefore, to improve the bolometer's responsivity, its electrical resistance and thermal impedance should be increased.

The bolometers were traditionally fabricated as miniature thermistors, suspended by tiny wires. Another popular method of bolometer fabrication is the use of metal film depositions [12, 13], usually of nichrome. In many modern bolometers, a thermoresistive, thin-film material is deposited on the surface of a micromachined silicon, or a glass membrane that is supported by a silicon frame. This approach gains popularity with the increased demand for the focal-plane array sensors (FPA) that are required for the thermal imaging. When an application does not need a high sensitivity and where the cost of fabrication is a critical factor, a platinum film bolometer is an attractive choice. Platinum has a small but predictive temperature coefficient of resistivity. The platinum film (having thickness of about 500 Å) may be deposited and photolithographically patterned over the thin glass membrane (Fig. 14.27a). The membrane is supported in the cavity etched in silicon by tiny extended leads. So, the membrane plate is virtually floating over the V-grooved cavity in the Si substrate. This helps to dramatically minimize its thermal coupling with the substrate (increase the thermal resistance). Figure 14.27b shows a microphotograph of an array of the PT bolometers used for the thermal imaging.

Besides platinum, many other materials may be used as temperature-sensitive resistors, for example, polysilicon, germanium, TaNO, and others. An important issue when selecting a particular material is its compatibility with a standard CMOS process so that a full monolithic device can be fabricated on a single silicon chip, including the interface electronic circuit. Thus, polysilicon is an attractive choice, along with deposition of germanium films (Fig. 14.28).

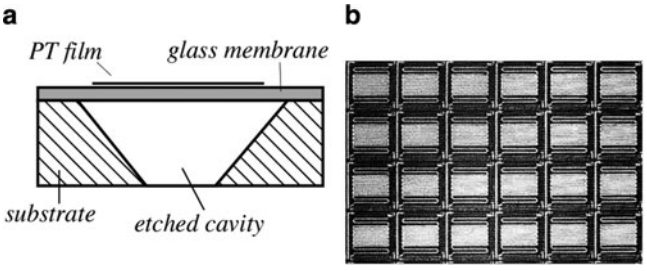

Fig. 14.27 Platinum film bolometer glass membrane over the etched cavity (**a**) and array of bolometers (**b**)

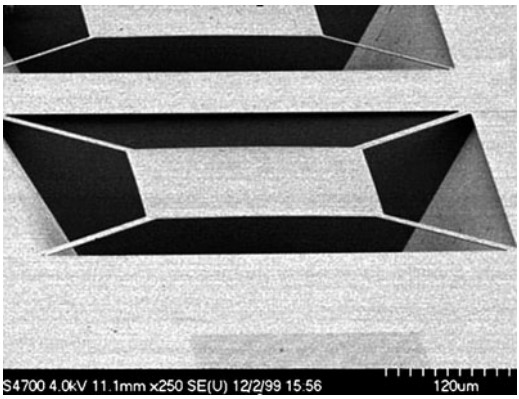

Fig. 14.28 Germanium film bolometer floating over the silicon cavity (courtesy of Prof. J. Shie)

As it follows from (14.24), one of the critical issues that always must be resolved when designing a bolometer (or any other accurate temperature sensor, for that matter) is to assure good thermal insulation of the sensing element from the supporting structure, connecting wires, and interface electronics. Otherwise, heat loss from the element may result in large errors and reduced sensitivity (see Sect. 16.1). One method to achieve this is to completely eliminate any metal conductors and to measure the temperature of the bolometer by using a fiber optic technique, as it has been implemented in the *E*-field probe fabricated by Luxtron, Mountain View, CA (U.S. Patent No. 4,816,634). In the design (Fig. 14.26b), a miniature bolometer is suspended in the end of an optical probe, and its temperature is measured by a fluoroptic® temperature sensor (see Sect. 16.6.1), while another similar optical sensor measures ambient temperature to calculate ΔT.

14.7.5 Active Far-Infrared Sensors

In the active far-infrared (AFIR) sensor, a process of measuring thermal radiation flux is different from previously described passive-infrared detectors (PIR). Contrary to a PIR sensing element, whose temperature depends on both the ambient and object's temperatures, the AFIR sensor's surface is actively controlled by a special circuit to have a defined temperature T_s that in most applications is maintained constant during an entire measurement process. To control the sensor's surface temperature, electric power P is provided by a control (or excitation) circuit (Fig. 14.29a). To regulate T_s, the circuit measures element's surface temperature and compares it with an internal reference.

Obviously, the incoming power maintains T_s higher than ambient. In some applications, T_s may be selected higher than the highest temperature of the object; however, in most cases, just several tenths of a degree Celsius above the ambient is sufficient. Since the element's temperature is above ambient, the sensing element loses thermal energy toward its surroundings, rather than passively absorbs it, as in

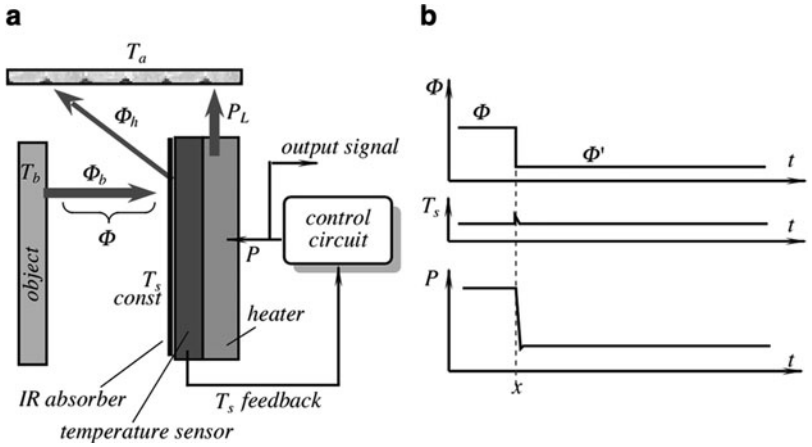

Fig. 14.29 AFIR element radiates thermal flux Φ_η toward its housing and absorbs flux Φ_b from the object (**a**); timing diagrams for radiative flux, surface temperature, and supplied power (**b**)

a PIR detector. Part of the heat loss is in the form of a thermal conduction, part is a thermal convection, and the other part is thermal radiation. That third part is the one that has to be measured. Unlike the conductive and convective heat transfer that is always directed out of the sensing element (because it is warmer than ambient), the radiative heat transfer may go in either direction, depending on temperature of the object. The radiative flux is governed by the fundamental (3.133), which is known as the Stefan–Boltzmann law.

Some of the radiation power goes out of the element to the sensor's housing, while the other is coming from the object (or goes to the object). What is essential is the net thermal flow (conductive + convective + radiative) always must come out of the sensor, e.g., it must have a negative sign.

If the AFIR element is provided with a cooling element (for instance, a thermoelectric device operating on Peltier effect[3]), T_s may be maintained at or below ambient. However, from the practical standpoint, it is easier to warm the element up rather than to cool it down. In the following, we discuss the AFIR sensors where the surface is warmed up either by an additional heating element or due to a self-heating effect in a temperature sensor [8, 14, 15, 16].

Dynamically, temperature T_s of any thermal element, either active or passive, in general terms may be described by the first-order differential equation

$$cm\frac{dT_s}{dt} = P - P_L - \Phi, \tag{14.25}$$

where P is the power supplied to the element from a power supply or an excitation circuit (if any), P_L is a nonradiative thermal loss that is attributed to thermal

[3]See Section 3.9.

conduction and convection, m and c are the sensor's mass and specific heat, respectively, and $\Phi = \Phi_n + \Phi_b$ is the net radiative thermal flux. We select a positive sign for power P when it is directed toward the element.

In a PIR detector, for instance, in the thermopile or pyroelectric, no external power is supplied ($P = 0$); hence, the speed response depends only on the sensor's thermal capacity and heat loss and is characterized by a thermal time constant τ_T. In the AFIR element, after a warm-up period, the control circuit forces the element's surface temperature T_s to stay constant, which means

$$\frac{dT_s}{dt} = 0, \tag{14.26}$$

and (14.25) becomes algebraic:

$$P = P_L + \Phi. \tag{14.27}$$

Contrary to the PIR sensors, the AFIR detector acts as an "infinite" heat source. It follows from the above that under the idealized conditions, its response does not depend on thermal mass and is not a function of time. If the control circuit is highly efficient, since P_L is constant at given ambient conditions, electronically supplied power P should track changes in the radiated flux Φ with high fidelity. A magnitude of that power may be used as the sensor's output signal. Equation (14.27) predicts that an AFIR element, at least in theory, is a much faster device if compared with the PIR. The efficiency of the AFIR detector is a function of both its design and the control circuit.

Nonradiative loss P_L is a function of the ambient temperature T_a and a loss factor α_s:

$$P_L = \alpha_s(T_s - T_a). \tag{14.28}$$

To generate heat in the AFIR sensor, it may be provided with a heating element having electrical resistance R. During the operation, electric power dissipated by the heating element is a function of voltage V across that resistance

$$P = V^2/R. \tag{14.29}$$

Let us assume that the AFIR sensor is used in a radiation thermometer. Its output signal should be representative of the object's temperature T_b that is to be measured. Substituting (3.138, 14.28, and 14.29) into (14.27), assuming that $T = T_b$ and $T_s > T_a$, after simple manipulations the object's temperature may be presented as a function of voltage V across the heating element:

$$T_b = \sqrt[4]{T_s^4 - \frac{1}{A\sigma\varepsilon_s\varepsilon_b}\left[\frac{V^2}{R} - \alpha_s(T_s - T_a)\right]}. \tag{14.30}$$

Coefficient α_s has a meaning of thermal conductivity from the AFIR-sensing element to the environment (housing).

Contrary to a PIR detector, an AFIR sensor is active and can generate a signal only when works in orchestra with the control circuit. The control circuit must include the following essential components: a reference to set the preset temperature, an error amplifier, and a driver stage. In addition, it may include an *RC* network for correcting a loop response function and for stabilizing its operation; otherwise, an entire system may be prone to oscillations [17].

It may be noted that an AFIR sensor along with its control circuit is a direct converter of thermal radiative power into electric voltage and quite efficient one. Its typical responsivity is in the range of 3,000 V/W, which is much higher when comparing with a thermopile whose typical responsivity is in the range of 100 V/W. An efficient way to fabricate an AFIR sensor would be by MEMS technology. In fact, an AFIR sensor is a close relative to a bolometer as described in the previous section. It just needs to be provided with a heater that can be deposited beneath the bolometer-sensing element.

14.8 Optical Design

An optical sensor usually operates in conjunction with some kind of an optical system that may be as simple as a window or spectral filter. Often, an optical system includes lenses, mirrors, fiber optics, and other elements that alter direction of light. In many applications, the sensor works together with a special light source whose properties must be matched to the sensor both spectrum-wise and space-wise.

Several factors are critical in designing the optical components for an optical sensor. Let us, for instance, take a point light source that should be detected by a photodetector (Fig. 14.30a). In most cases, the sensor's output signal is proportional to the received photonic power, which, in turn, proportional to the sensor's surface

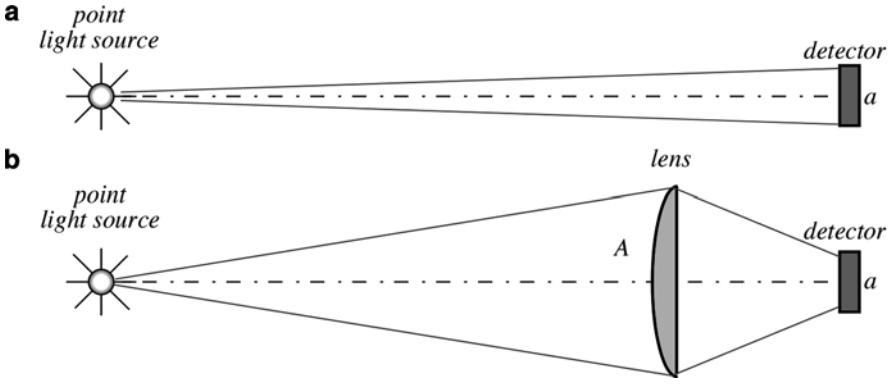

Fig. 14.30 Efficiency of a detector depends on (**a**) its surface area *a* and (**b**) the input aperture *A* of the focusing system

area (input aperture). Yet, the sensing area is usually small. For example, a thermopile pixel of the sensor shown in Fig. 14.21a has a sensing area 0.022 mm^2, and thus it will receive a very small portion of the photon flux emanated from the object. Figure 14.30b shows that the use of a focusing lens can dramatically increase the received light flux. This is because the entire lens aperture receives the flux and, like a funnel, brings it to a small sensor. Efficiency of a single lens depends on its refractive index n (see Chap. 4). The overall improvement in the sensitivity can be estimated by use of (4.5) and (4.8):

$$k \approx \frac{A}{a}\left[1 - 2\left(\frac{n-1}{n+1}\right)^2\right], \tag{14.31}$$

where A and a are respective effective areas (apertures) of the lens and the sensing area of a photodetector.

For glasses and most plastics operating in the visible- and near-infrared spectral ranges, the equation can be simplified to

$$k \approx 0.92\frac{A}{a}. \tag{14.32}$$

Thus, the amount of light received by the sensing element is proportional to the lens aperture area. That is why good photographic lenses have quite large apertures. It should be pointed out that arbitrary placement of a lens may be more harmful than helpful. That is, a lens system must be carefully planed to be effective. For instance, many photodetectors have built-in lenses that are effective for parallel rays. If an additional lens is introduced in front of such a detector, it will create nonparallel rays at the input resulting in misalignment of the optical system and poor performance. Thus, whenever additional optical devices need to be employed, detector's own optical properties must be considered.

14.9 Gas Flame Detectors

Detection of a gas flame is very important for security and fire prevention systems. In many respects, it is a more sensitive way to detect fire than a smoke detector, especially outdoors where smoke concentration may not reach a threshold level for the alarm triggering.

To detect burning gas, it is possible to use a unique feature of the flame, a noticeable portion of its optical spectrum is located in the ultraviolet (UV) spectral range (Fig. 14.31). Sunlight, after passing through the atmosphere loses a large portion of its UV spectrum located below 250 nm, while a gas flame contains UV components down to 180 nm. This makes it possible to design a narrow-bandwidth element for the UV spectral range that is selectively sensitive to flame and not sensitive to the sunlight or electric lights.

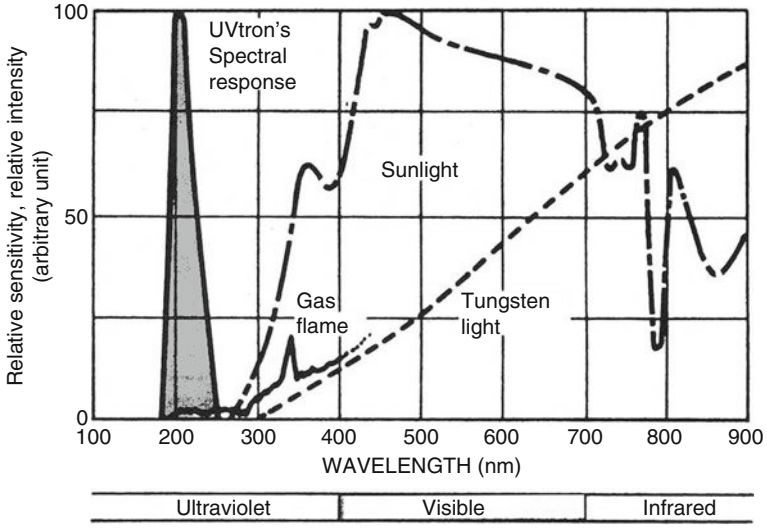

Fig. 14.31 Electromagnetic spectra of various sources (courtesy of Hamamatsu Photonics K.K.)

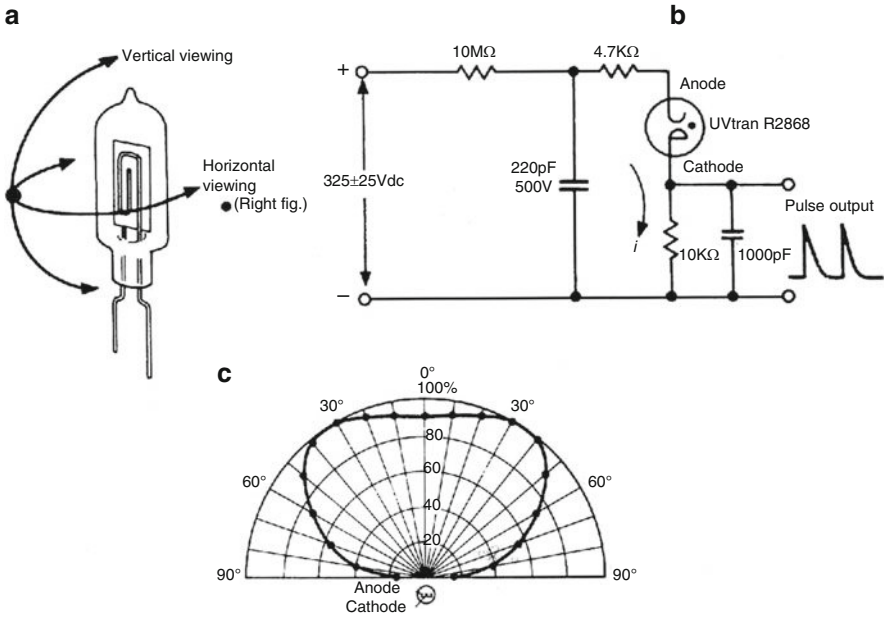

Fig. 14.32 UV flame detector glass-filled tube (**a**); angle of view in horizontal plane (**b**); and recommended operating circuit (**c**) (courtesy of Hamamatsu Photonics K.K.)

An example of such a device is shown in Fig. 14.32a. The element is a UV detector that makes use of a photoelectric effect in metals along with the gas multiplication effect (see Chap. 15). The detector is a rare-gas-filled tube. The UV-transparent housing assures wide angles of view in both horizontal and vertical planes (Fig. 14.32c). The device needs high voltage for operation and under normal conditions is not electrically conductive. Upon being exposed to a flame, the high-energy UV photons strike the cathode-releasing free electrons to the gas-filled tube interior. Gas atoms receive an energy burst from the emitted electrons, which results in gas luminescence in the UV spectral range. This, in turn, cause more electrons to be emitted, which cause more UV luminescence. Thus, the element develops a fast avalanche-type electron multiplication making the anode–cathode region electrically conductive. Hence, upon being exposed to a gas flame, the element works as a current switch producing a strong positive voltage spike at its output (Fig. 14.30b). It follows from the above description that the element generates UV radiation in response to the flame detection. Albeit being of a low intensity, the UV does not present harm to people; however, it may lead to cross-talk between the similar neighboring sensors.

References

1. Chappell A (ed) (1978) Optoelectronics: theory and practice. McGraw Hill, New York
2. Spillman WB, Jr (1991) Optical detectors. In: Udd Eric (ed) Fiber optic sensors. Wiley, New York pp. 69–97
3. Verdeyen JT (1981) Laser electronics. Prentice-Hall, Englewood Cliffs, NJ
4. Graeme J (1992) Phase compensation optimizes photodiode bandwidth. EDN May 7, pp. 177–183
5. Völklein A, Wiegand A, Baier V (1991) Sens Actuators A 29:87–91
6. Schieferdecker J, Quad R, Holzenkämpfer E, Schulze M (1995) Infrared thermopile sensors with high sensitivity and very low temperature coefficient. Sens Actuators A, 46–47, 422–427
7. Meixner H, Mader G, Kleinschmidt P (1986) Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). Siemens Forsch-u Entwicl Ber Bd 15(3):105–114
8. Fraden J (1991) Noncontact temperature measurements in medicine. Chapter 17, In: Wise D (ed) Bioinstrumentation and biosensors. Marcel Dekker, New York, pp 511–549
9. Fraden J (1989) Infrared electronic thermometer and method for measuring temperature. US Patent 4,797,840, 10 Jan
10. Fraden J (1988) Motion detector. US Patent 4,769,545, 6 Sept
11. Astheimer RW (1984) Thermistor infrared detectors. SPIE No. 443, pp 95–109
12. Shie, J-S, Weng PK (1991)Fabrication of micro-bolometer on silicon substrate by anizotropic etching technique. In: Transducers'91. International Conference on Solid-state Sensors and Actuators. Digest of Technical Papers. ©IEEE, pp 627–630
13. Vogl TP, Shifrin GA, Leon BJ (1962) Generalized theory of metal-film bolometers. J Opt Soc Am 52:957–964
14. Fraden J (1992) Active far infrared detectors. In: Schooley JF (ed) Temperature. Its measurement and control in science and industry, vol. 6, part 2, ©American Institute of Physics, pp 831–836
15. Fraden J (1989) Radiation thermometer and method for measuring temperature. US Patent 4,854,730 8 Aug

16. Fraden J (1990) Active infrared motion detector and method for detecting movement US Patent 4,896,039, 23 Jan
17. Mastrangelo CH, Muller RS (1991) Design and performance of constant-temperature circuits for microbridge-sensor applications. In: Transducers'91. International Conference on Solid-state Sensors and Actuators. Digest of TECHNICAL Papers, IEEE pp 471–474

Chapter 15
Radiation Detectors

To understand something,
means to derive it from quantum mechanics,
which nobody understands.

– Joe Fineman

Figure 3.41 shows a spectrum of the electromagnetic waves. On its left-hand side, there is a region of the γ-radiation. Then, there are the X-rays that depending on the wavelengths are divided into hard, soft, and ultrasoft rays. However, a spontaneous radiation from the matter not necessarily should be electromagnetic: There is the so-called nuclear radiation, which is emission of particles from the atomic nuclei. A spontaneous decay can be of two types: The charged particles (α and β particles, and protons) and uncharged particles that are the neutrons. Some particles are complex like the α particles, which are the nuclei of helium atoms consisting of two neutrons, while other particles are generally simpler, like the β particles that are either electrons or positrons. Ionizing radiations are given that name because as they pass through various media that absorb their energy, additional ions, photons, or free radicals are created.

Certain naturally occurring elements are not stable but slowly decompose by throwing away a portion of their nucleus. This is called radioactivity. It was discovered in 1896 by Henry Becquerel while working on phosphorescent materials. These materials glow in the dark after exposure to light, and he thought that the glow produced in cathode ray tubes by X-rays might be connected with phosphorescence. He wrapped a photographic plate in black paper and placed various phosphorescent minerals on it. All results were negative until he used uranium ($Z = 92$)[1] salts. The result with these compounds was a deep blackening of the plate. These radiations were called Becquerel rays.

Besides the naturally occurring radioactivity, there are many man-made nuclei that are radioactive. These nuclei are produced in nuclear reactors, which may yield

[1] Z is the atomic number.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_15, © Springer Science+Business Media, LLC 2010

highly unstable elements. The other source of radiation is the space from which the earth is constantly bombarded by the particles.

Regardless of the sources or ages of radioactive substances, they decay in accordance with the same mathematical law. The law is stated in terms of number N of nuclei still undecayed, and dN the number of nuclei that decay in a small interval dt. It was proven experimentally that

$$\mathrm{d}N = -\lambda N \mathrm{d}t, \tag{15.1}$$

where λ is a decay constant specific for a given substance. From (15.1), it can be defined as the fraction of nuclei that decay in unit time

$$\lambda = -\frac{1}{N}\frac{\mathrm{d}N}{\mathrm{d}t}. \tag{15.2}$$

The SI unit of radioactivity is the *becquerel* (Bq), which is equal to the activity of radionuclide decaying at the rate of one spontaneous transition per second. Thus, the becquerel is expressed in a unit of time: Bq $= \mathrm{s}^{-1}$. To convert to the old historical unit, which is the *curie*, the becquerel should be multiplied by 3.7×10^{10} (Table A.4).

The absorbed dose is measured in *grays* (Gy). A gray is the absorbed dose when the energy per unit mass imparted to matter by ionizing radiation is 1 J/kg. That is, Gy $=$ J/kg.

When it is required to measure exposure to X- and γ-rays, the dose of ionizing radiation is expressed in coulombs per kg, which is an exposure resulting in the production of 1 C of electric charge per 1 kg of dry air. In SI, a unit of C/kg replaces an older unit of *roentgen*.

The function of any radiation detector depends on the manner in which the radiation interacts with the material of the detector itself. There are many excellent texts available on the subject of detecting radioactivity, for instance [1, 2].

There are four general types of radiation detectors: the scintillation detectors, the gaseous detectors, the liquid detectors, and the semiconductor detectors. Further, all detectors can be divided into two groups according to their functionality: the collision detectors and the energy detectors. The former merely detect the presence of a radioactive particle, while the latter can measure the radiative energy. That is, all detectors can be either quantitative or qualitative.

15.1 Scintillating Detectors

The operating principle of these detectors is based on the ability of certain materials to convert nuclear radiation into light. Thus, an optical photon detector in a combination with a scintillating material can form a radiation detector. It should be noted, however, that despite of a high efficiency of the conversion, the light intensity resulting from the radiation is extremely small. This demands photomultipliers to magnify signals to a detectable level.

The ideal scintillation material should posses the following properties:

1. It should convert the kinetic energy of charged particles into detectable light with a high efficiency.
2. The conversion should be linear. That is, the light produced should be proportional to the input energy over a wide dynamic range.
3. The postluminescence (the light decay time) should be short to allow fast detection.
4. The index of refraction of the material should be near that of glass to allow efficient optical coupling of the light to the photomultiplier tube.

The most widely used scintillators include the inorganic alkali halide crystals (of which sodium iodine is the favorite) and organic-based liquids and plastics. The inorganics are more sensitive, but generally slow, while organics are faster, but yield less light.

One of the major limitations of scintillation counters is their relatively poor energy resolution. The sequence of events that leads to the detection involves many inefficient steps. Therefore, the energy required to produce one information carrier (a photoelectron) is in the order of 1,000 eV or more, and the number of carriers created in a typical radiation interaction is usually no more than a few thousand. For example, the energy resolution for sodium iodine scintillators is limited to about 6% when detecting 0.662 MeV γ-rays and is largely determined by the photoelectron statistical fluctuations. The only known way to reduce the statistical limit on energy resolution is to increase the number of information carriers per pulse. This can be accomplished by the use of semiconductor detectors that are described in the later paragraphs.

A general simplified arrangement of a scintillating sensor is shown in Fig. 15.1 in conjunction with a photomultiplier. The scintillator is attached to the front end of the photomultiplier (PM). The front end contains a photocathode that is maintained at a ground potential. There is a large number of special plates called dynodes positioned inside the PM tube in an alternating pattern, reminding a shape of a

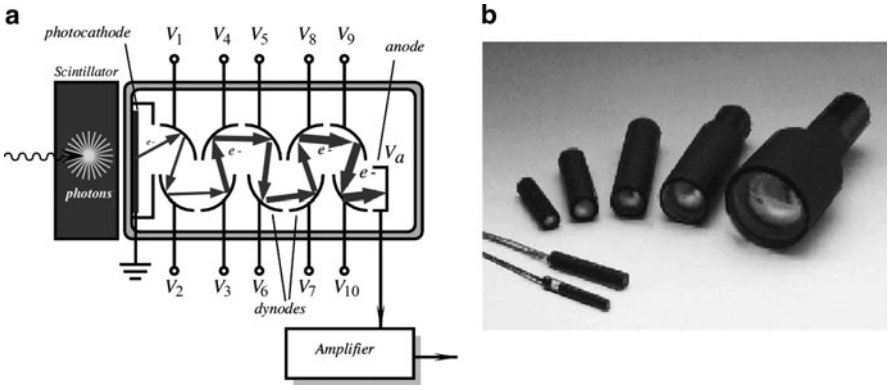

Fig. 15.1 Scintillation detector with a photomultiplier

"venetian blind." Each dynode is attached to a positive voltage source in a manner that the farther the dynode from the photocathode, the higher is its positive potential. The last component in the tube is an anode, which has the highest positive potential, sometimes on the order of several thousand volts. All components of the PM are enveloped into a glass vacuum tube that may contain some additional elements like focusing electrodes, shields, etc.

Although the PM is called a photomultiplier, in reality it is an electron multiplier, as there are no photons, only electrons are inside the PM tube during its operation. For the illustration, let us assume that a γ-ray particle has a kinetic energy of 0.5 MeV (megaelectron volt). It is deposited on the scintillating crystal resulting in a number of liberated photons. In thallium-activated sodium iodine, the scintillating efficiency is about 13%; therefore, total of $0.5 \times 0.13 = 0.065$ MeV or 65 keV of energy is converted into visible light with an average energy of 4 eV. Therefore, about 15,000 scintillating photon s are produced per gamma pulse. This number is too small to be detected by an ordinary photodetector; hence, a multiplication effect is required before the actual detection takes place. Of 15,000 photons, probably about 10,000 reach the photocathode, whose quantum efficiency is about 20%. The photocathode serves to convert incident light photons into low-energy electrons. Therefore, the photocathode produces about 2,000 electrons per pulse. The PM tube is a linear device, that is, its gain is almost independent of the number of multiplied electrons.

Since all dynodes are at positive potentials (V_1 to V_{10}), an electron released from the photocathode is attracted to the first dynode, liberating several very-low-energy electrons at impact with its surface. Thus, a multiplication effect takes place at the dynode. These electrons will be easily guided by the electrostatic field from the first to the second dynode. They strike the second dynode and produce more electrons that travel to the third dynode and so on. The process results in an increasing number of available electrons (avalanche effect). An overall multiplication ability of a PM tube is in the order of 10^6. As a result, about 2×10^9 electrons will be available at a high-voltage anode (V_a) for the production of electric current. This is a pretty strong electric current that can be easily processed by an electronic circuit. A gain of a PM tube is defined as

$$\mathrm{G} = \alpha\delta^N, \tag{15.3}$$

where N is the number of dynodes, α is the fraction of electrons collected by the PM tube, δ is the efficiency of the dynode material, that is, the number of electrons liberated at impact. Its value ranges from 5 to 55 for a high-yield dynode. The gain is sensitive to the applied high voltage, because δ is almost a linear function of the interdynode voltage.

A modern design of a photomultiplier is called the channel photomultiplier or CPM for short use. It is the evolution of the classical photomultiplier tube PM. The modern CPM technology preserves the advantages of the classical PM while avoiding its disadvantages. Figure 15.2a shows a faceplate with a photocathode, the bent channel amplification structure, and the anode. Like the PM of Fig. 15.1, photons in the CPM are converted inside the photocathode into photoelectrons and accelerated

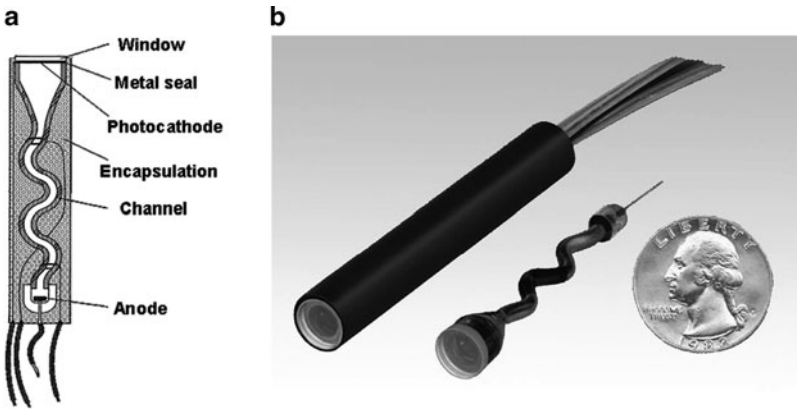

Fig. 15.2 Channel photomultiplier: Cross-sectional view (**a**) and external view with a potted encapsulation at left and with no encapsulation at right (**b**) (Courtesy of PerkinElmer, Inc.)

in a vacuum toward the anode by an electrical field. Instead of the complicated dynode structure, there is a bent, thin semiconductive channel, which the electrons have to pass. Each time when electrons hit the wall of the channel, secondary electrons are emitted from the surface. At each collision, there is a multiplication of the secondary electrons, resulting in an avalanche effect. Ultimately, an electron multiplication of 10^9 and more can be obtained. The resulting current can be read out at the anode. The CPM detector is potted with encapsulation material and is quite rugged when compared to the fragile PM. Magnetic field disturbance is negligibly small. Figure 15.2a shows pictures of the CPM. An important advantage of the CPM technology is its very low background noise. The term background noise refers to the measured output signal in the absence of any incident light. With classical PMs, the background noise originating from the dynode structure is generally a non-negligible part of the total background. As a result, the only effective source of background for the CPM is generated from the thermal emission of the photocathode. Since the CPM is manufactured in a monolithic semiconductive channel structure, no charge-up effects can occur as known of classical PMs with isolating glass bulbs. Hence, extremely stable background conditions are observed. No sudden bursts occur. Also, due to the absence of dynode noise, a very clean separation between an event created from a photoelectron and electronic noise can be performed. This leads to high stability of the signal over time.

15.2 Ionization Detectors

These detectors rely on the ability of some gaseous and solid materials to produce ion pairs in response to the ionization radiation. Then, positive and negative ions can be separated in an electrostatic field and measured.

Ionization happens because charged particles upon passing at a high velocity through an atom can produce sufficient electromagnetic forces, resulting in the separation of electrons, thus creating ions. Remarkably, the same particle can produce multiple ion pairs before its energy is expended. Uncharged particles (like neutrons) can produce ion pairs at collision with the nuclei.

15.2.1 Ionization Chambers

These radiation detectors are the oldest and most widely used. The ionizing particle causes the ionization and excitation of gas molecules along its passing track. As a minimum, the particle must transfer an amount of energy equal to the ionization energy of the gas molecule to permit the ionization process to occur. In most gasses of interest for radiation detection, the ionization energy for the least tightly bound electron shells is between 10 and 20 eV [2]. However, there are other mechanisms by which the incident particle may lose energy within gas that do not create ions, for instance, moving gas electrons to a higher energy level without removing it. Therefore, the average energy lost by a particle per ion pair formed (called W value) is always greater than the ionizing energy. The W value depends on gas (Table 15.1), the type of radiation, and its energy.

In the presence of an electric field, the drift of the positive and negative charges represented by the ions end electrons constitutes an electric current. In a given volume of gas, the rate of the formation of the ion pair is constant. For any small volume of gas, the rate of formation will be exactly balanced by the rate at which ion pairs are lost from the volume, either through recombination, or by diffusion or migration from the volume. If recombination is negligible and all charges are effectively collected, the steady-state current produced is an accurate measure of the rate of ion pair formation. Figure 15.3a illustrates a basic structure of an ionizing chamber and the current/voltage characteristic.

A volume of gas is enclosed between the electrodes, which produce an electric field. An electric current meter is attached in series with the voltage source E and the electrodes. There is no electrical conduction and no current under the no-ionization conditions. Incoming radiation produces, in the gas, positive and negative ions that are pulled by the electric field toward the corresponding electrodes forming an electric current. The current vs. voltage characteristic of the chamber is

Table 15.1 W values for different gases (adapted from [2])

Gas	W value (in eV/ion pair)	
	Fast electrons	Alphas
A	27.0	25.9
He	32.5	31.7
N_2	35.8	36.0
Air	35.0	35.2
CH_4	30.2	29.0

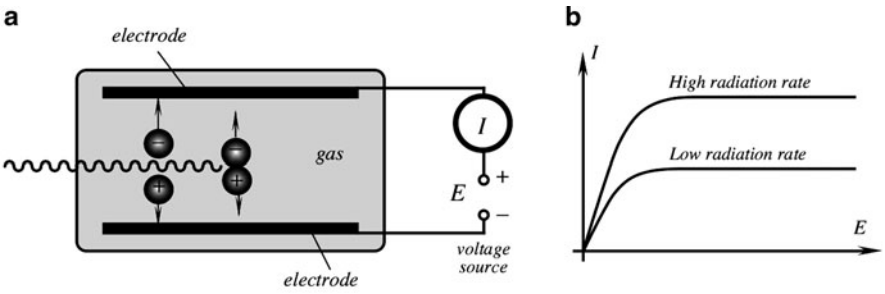

Fig. 15.3 Simplified schematic of an ionization chamber (**a**) and a current vs. voltage characteristic (**b**)

shown in Fig. 15.3b. At relatively low voltages, the ion recombination rate is strong and the output current is proportional to the applied voltage, because higher voltage reduces the number of recombined ions. A sufficiently strong voltage completely suppresses all recombinations by pulling all available ions toward the electrodes, and the current becomes voltage-independent. However, it still depends on the intensity of irradiation. This region is called saturation and where the ionization chamber normally operates.

15.2.2 Proportional Chambers

The proportional chamber is a type of a gas-filled detector that almost always operates in a pulse mode and relies on the phenomenon of a gas multiplication. This is why these chambers are called the proportional counters. Due to gas multiplication, the output pulses are much stronger than in conventional ion chambers. These counters are generally employed in the detection and spectroscopy of low-energy X radiation and for the detection of neutrons. Contrary to the ionization chambers, the proportional counters operate at higher electric fields, which can greatly accelerate electrons liberated during the collision. If these electrons gain sufficient energy, they may ionize a neutral gas molecule, thus creating an additional ion pair. Hence, the process is of an avalanche type resulting in a substantial increase in the electrode current. The name for this process is Townsend avalanche. In the proportional counter, the avalanche process ends when the electron collides with the anode. Since the electron must reach the gas ionization level in the proportional counter, there is a threshold voltage after which the avalanche process occurs. In typical gases at atmospheric pressure, the threshold field level is on the order of 10^6 V/m.

Differences between various gas counters are illustrated in Fig. 15.4. At very low voltages, the field is insufficient to prevent the recombination of ion pairs. In the saturation level, all ions are drifted to the electrodes. A further increase in voltage results in gas multiplication. Over some region of the electric field, the gas

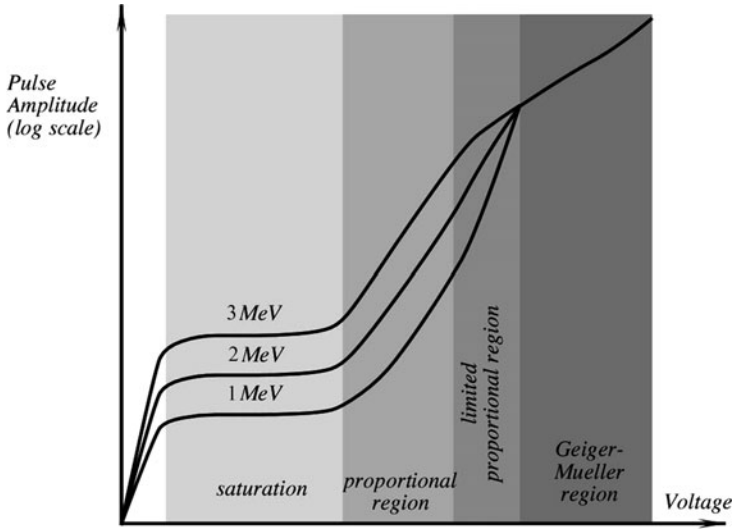

Fig. 15.4 Various operating voltages for gas-filled detectors (adapted from [2])

multiplication will be linear, and the collected charge will be proportional to the number of original ion pairs created during the ionization collision. An even further increase in the applied voltage can introduce nonlinear effects, which are related to the positive ions due to their slow velocity.

15.2.3 Geiger–Müller Counters

The Geiger–Müller (GM) counter was invented in 1928[2] and is still in use thanks to its simplicity, low cost, and ease of operation. The GM counter is different from other ion chambers by its much higher applied voltage (see Fig. 15.4). In the region of the GM operation, the output pulse amplitude does not depend on the energy of ionizing radiation and is strictly a function of the applied voltage. A GM counter is usually fabricated in the form of a tube with an anode wire in the center (Fig. 15.5). The tube is filled with a noble gas such as helium or argon. A secondary component

[2]Johannes (Hans) Wilhelm Geiger (1882–1945) was a German physicist. He is best known as the coinventor of the GM counter and for the Geiger–Marsden experiment that discovered the atomic nucleus. He was a "loyal Nazi" who unhesitatingly betrayed many of his former colleagues. Walther Müller (1905–1979) was a student of Geiger and a founder of a U.S. company that produced tubes for the GM counters.

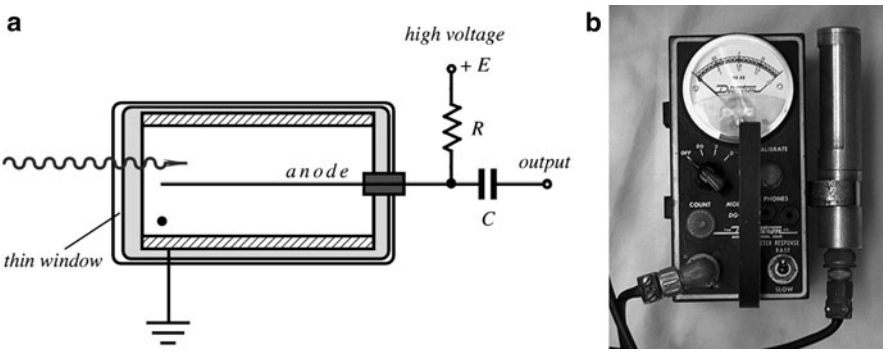

Fig. 15.5 Circuit of a Geiger–Müller counter, *filled circle* indicates gas (**a**) and a vintage GM counter (**b**)

is usually added to the gas for the purpose of quenching, which is preventing of a retriggering of the counter after the detection. The retriggering may cause multiple pulses instead of the desired one. The quenching can be accomplished by several methods; some of them are short-time reduction in the high voltage applied to the tube, use of high-impedance resistors in series with the anode, and adding the quench gas at concentrations of 5–10%. Many organic molecules possess the proper characteristics to serve as a quench gas. Of these, ethyl alcohol and ethyl formate have proven to be the most popular.

In a typical avalanche created by a single original electron, the secondary ions are created. In addition to them, many excited gas molecules are formed. Within a few nanoseconds, these excited molecules return to their original state through the emission of energy in the form of ultraviolet (UV) photons. These photons play an important role in the chain reaction occurring in the GM counter. When one of the UV photons interacts by photoelectric absorption in some other region of the gas, or at the cathode surface, a new electron is liberated that can subsequently migrate toward the anode and will trigger another avalanche. In a Geiger discharge, the rapid propagation of the chain reaction leads to many avalanches that initiate at random radial and axial positions throughout the tube. Secondary ions are therefore formed throughout the cylindrical multiplying region that surrounds the anode wire. Hence, the discharge grows to envelope the entire anode wire, regardless of the position at which the primary initiating event occurred.

Once the Geiger discharge reaches a certain level, however, collective effects of all individual avalanches come into play and ultimately terminate the chain reaction. This point depends on the number of avalanches and not on the energy of the initiating particle. Thus, the GM current pulse is always of the same amplitude, which makes the GM counter just an indicator of irradiation, because all information on the ionizing energy is lost.

In the GM counter, a single particle of a sufficient energy can create about 10^9 to10^{10} ion pairs. Because a single ion pair formed within the gas of the GM counter

can trigger a full Geiger discharge, the counting efficiency for any charged particle that enters the tube is essentially 100%. However, the GM counters are seldom used for counting neutrons because of a very low efficiency of counting. The efficiency of GM counters for γ-rays is higher for those tubes constructed with a cathode wall of high-Z material. For instance, bismuth ($Z = 83$) cathodes have been widely used for the γ-detection in conjunction with gases of high atomic numbers, such as xenon and krypton, which yield a counting efficiency up to 100% for photon energies below about 10 keV.

A further improvement of GM counter is the so-called *wire chamber* that contains many parallel wires, arranged as a grid. A high voltage is applied to the wires with the metal casing being at a ground potential. As in the GM counter, a particle leaves a trace of ions and electrons, which drift toward the case or the nearest wire, respectively. By marking off the wires that had a pulse of current, one can see the particle's path.

15.2.4 Semiconductor Detectors

The best energy resolution in modern radiation detectors can be achieved in the semiconductor materials, where comparatively a large number of carriers for a given incident radiation event occurs. In these materials, the basic information carriers are electron–hole pairs created along the path taken by the charged particle through the detector. The charged particle can be either primary radiation or a secondary particle. The electron–hole pairs in some respects are analogous to the ion pairs produced in the gas-filled detectors. When an external electric field is applied to the semiconductive material, the created carriers form a measurable electric current. The detectors operating on this principle are called solid-state or semiconductor diode detectors. The operating principle of these radiation detectors is the same as that of the semiconductor light detectors. It is based on the transition of electrons from one energy level to another when they gain or lose energy. For the introduction to the energy band structure in solids, the reader should refer to Sect. 14.1.

When a charged particle passes through a semiconductor with the band structure shown in Fig. 14.1, the overall significant effect is the production of many electron–hole pairs along the track of the particle. The production process may be either direct or indirect, in that the particle produces high-energy electrons (or Δ rays) that subsequently lose their energy in the production of more electron–hole pairs. Regardless of the actual mechanism involved, what is of interest to our subject is the average energy expended by the primary charged particle produces one electron–hole pair. This quantity is often called the "ionization energy." The major advantage of semiconductor detectors lies in the smallness of the ionization energy. The value of it for silicon or germanium is about 3 eV, compared with 30 eV required to create an ion pair in typical gas-filled detectors. Thus, the number of charge carriers is about ten times greater for the solid-state detectors for a given energy of a measured radiation.

To fabricate a solid-state detector, at least two contacts must be formed across a semiconductor material. For detection, the contacts are connected to the voltage source, which enables carrier movement. The use of a homogeneous Ge or Si, however, would be totally impractical. The reason for that is in an excessively high-leakage current caused by the material's relatively low resistivity (50 kΩ cm for silicon). The external voltage, when applied to the terminals of such a detector may cause a current that is 3–5 orders of magnitude greater than a minute radiation-induced electric current. Thus, the detectors are fabricated with the blocking junctions, which are reverse biased to dramatically reduce leakage current. In effect, the detector is a semiconductor diode, which readily conducts (has low resistivity) when its anode (p side of a junction) is connected to a positive terminal of a voltage source and the cathode (an n side of the junction) to the negative. The diode conducts very little (it has very high resistivity) when the connection is reversed; thus, the name reverse biasing is implied. If the reverse bias voltage is made very large, in excess of the manufacturer specified limit, the reverse leakage current abruptly increases (the breakdown effect), which often may lead to a catastrophic deterioration of detecting properties or to the device destruction.

Several configurations of silicon diodes are currently produced. Some of them are diffused junction diodes, surface barrier diodes, ion-implanted detectors, epitaxial layer detectors, and others. The diffused junction and surface barrier detectors find widespread applications for the detection of α particles and other short-range radiation. A good solid-state radiation detector should possess the following properties:

1. Excellent charge transport
2. Linearity between the energy of the incident radiation and number of electron–hole pairs
3. Absence of free charges (low-leakage current)
4. Production of a maximum number of electron–hole pairs per unit of radiation
5. High detection efficiency
6. Fast response speed
7. Large collection area
8. Low cost

When using semiconductor detectors, several factors should be seriously considered. Among them are the dead band layer of the detector and the possible radiation damage. If heavy charged particles or other weakly penetrating radiations enter the detector, there may be a significant energy loss before the particle reaches the active volume of the semiconductor. The energy can be lost in the metallic electrode and in a relatively thick silicon body immediately beneath the electrode. This thickness must be measured directly by the user if an accurate compensation is desirable. The simplest and most frequently used technique is to vary the angle of incidence of a monoenergetic, charged particle radiation [2]. When the angle of incidence is zero (that is, perpendicular to the detector's surface), the energy loss in the dead layer is given by

$$\Delta E_0 = \frac{\mathrm{d}E_0}{\mathrm{d}x} t, \tag{15.4}$$

where t is the thickness of the dead layer. The energy loss for an angle of incidence of Θ is

$$\Delta E(\theta) = \frac{\Delta E_0}{\cos \Theta}. \tag{15.5}$$

Therefore, the difference between the measured pulse height for angles of incidence of zero and the Θ is given by

$$E' = [E_0 - \Delta E_0] - [E_0 - \Delta E(\Theta)] = \Delta E_0 \left(\frac{1}{\cos \Theta} - 1 \right). \tag{15.6}$$

If a series of measurements is made as the angle of incidence is varied, a plot of E' as a function of $(1/\cos\Theta)-1$ should be a straight line whose slope is equal to ΔE_0. Using tabular data for $\mathrm{d}E_0/\mathrm{d}x$ for the incident radiation, the dead layer thickness can be calculated from (15.4).

Any excessive use of the detectors may lead to some damage to the lattice of the crystalline structure, due to disruptive effects of the radiation being measured as it passes through the crystal. These effects tend to be relatively minor for lightly ionizing radiation (β particles or γ-rays) but can become quite significant under the typical conditions of use for heavy particles. For example, prolonged exposure of silicon surface barrier detectors to fusion fragments will lead to a measurable increase in leakage current and a significant loss in energy resolution of the detector. With extreme radiation damage, multiple peaks may appear in the pulse height spectrum recorded for monoenergetic particles.

Earlier mentioned diffused junction diodes and surface barrier diodes are not quite suitable for the detection of penetrating radiation. The major limitation is in the shallow active volume of these sensors, which rarely can exceed 2–3 mm. This is not nearly enough, for instance, for a γ-ray spectroscopy. A practical method to make detectors for a more penetrating radiation is the so-called ion-drifting process. The approach consists of creating a thick region with a balanced number of donor impurities, which add either p or n properties to the material. Under ideal conditions, when the balance is perfect, the bulk material would resemble the pure (intrinsic) semiconductor without either property. However, in reality, the perfect pn balance never can be achieved. In Si or Ge, the pure material with the highest possible purity tends to be of p-type. To accomplish the desired compensation, the donor atoms must be added. The most practical compensation donor is lithium. The fabrication process involves a diffusing of lithium through the p crystal so that the lithium donors greatly outnumber the original acceptors, creating a n-type region near the exposed surface. Then, temperature is elevated, and the junction is reverse biased. This results in a slow drifting of lithium donors into the p-type for

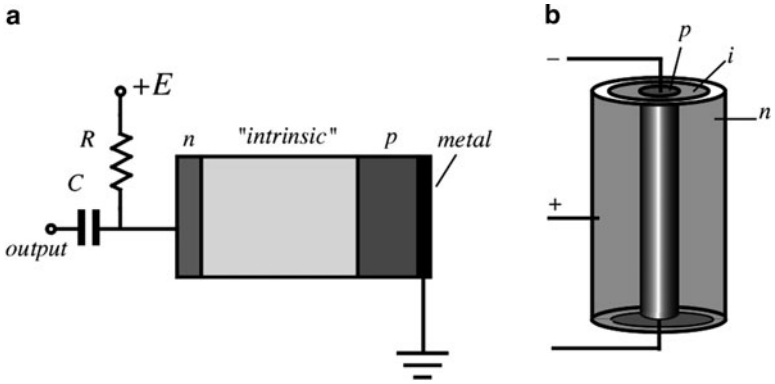

Fig. 15.6 Lithium-drifted pin-junction detector. Structure of the detector (**a**) and coaxial configuration of the detector (**b**)

the near perfect compensation of the original impurity. The process may take as long as several weeks. To preserve the achieved balance, the detector must be maintained at low temperature: 77 K for the germanium detectors. Silicon has very low ion mobility; thus, the detector can be stored and operated at room temperature. However, the lower atomic number for silicon ($Z = 14$) when compared with germanium ($Z = 32$) means that the efficiency of silicon for the detection of γ-rays is very low, and it is not widely used in the general γ-ray spectroscopy.

A simplified schematic of a lithium-drifted detector is shown in Fig. 15.6a. It consists of three regions where the "intrinsic" crystal is in the middle. In order to create detectors of a larger active volume, the shape can be formed as a cylinder (Fig. 15.6b), where the active volumes of Ge up to 150 cm^3 can be realized. The germanium, lithium-drifted detectors are designated as Ge(Li).

Table 15.2 Detecting the properties of some semiconductive materials (adapted from [2])

Material (operating temperature in K)	Z	Band gap (eV)	Energy per electron–hole pair (eV)
Si (300)	14	1.12	3.61
Ge (77)	32	0.74	2.98
CdTe (300)	48–52	1.47	4.43
HgI_2 (300)	80–53	2.13	6.5
GaAs (300)	31–33	1.43	4.2

Regardless of the widespread popularity of the silicon and germanium detectors, they are not the ideal from certain standpoints. For instance, germanium must always be operated at cryogenic temperatures to reduce thermally generated leakage current, while silicon is not efficient for the detection of γ-rays. There are some other semiconductors that are quite useful for the detection of radiation at room temperatures. Some of them are cadmium telluride (CdTe), mercuric iodine (HgI_2),

gallium arsenide (GaAs), bismuth trisulfide (Bi_2S_3), and gallium selenide (GaSe). Useful radiation detectors properties of some semiconductive materials are given in Table 15.2.

Probably the most popular at the time of this writing is cadmium telluride, which combines a relatively high-*Z* value (48 and 52) with band gap energy large enough (1.47 eV) to permit room temperature operation. Crystals of high purity can be grown from CdTe to fabricate the intrinsic detector. Alternatively, chlorine doping is occasionally used to compensate for the excess of acceptors and to make the material of a near-intrinsic type. Commercially available CdTe detectors range in size from 1 to 50 mm in diameter and can be routinely operated at temperatures up to 50°C without an excessive increase in noise. Thus, there are two types of CdTe detectors available: The pure intrinsic type and the doped type. The former has high-volume resistivity up to 10^{10} Ω cm, however, its energy resolution is not that great. The doped type has significantly better energy resolution, however, its lower resistivity (10^8 Ω cm) leads to a higher leakage current. Besides, these detectors are prone to polarization that may significantly degrade their performance.

In the solid-state detectors, it is also possible to achieve a multiplication effect as in the gas-filled detectors. An analog of a proportional detector is called an avalanche detector, which is useful for the monitoring of low-energy radiation. The gain of such a detector is usually in the range of several hundreds. It is achieved by creating within high-level semiconductor electric fields. Also, the radiation PSDs are available whose operating principle is analogous to the similar sensors functioning in the near-infrared region (see Sect. 7.4.6).

15.3 Cloud and Bubble Chambers

The cloud chamber, also known as the Wilson chamber,[3] is used for detecting particles of ionizing radiation. In its most basic form, a cloud chamber is a sealed environment containing a supercooled, supersaturated water or alcohol (e.g., methylated spirits) vapor, which is at the point of condensation. When an alpha particle or beta particle interacts with the mixture, it ionizes it. The resulting ions act as condensation nuclei, around which a mist will form. In other words, the vapor condenses into droplets when disturbed and ionized by the passage of a particle. A trail is left along the particle path, because many ions are being produced along the path of the charged particle. These tracks have distinctive shapes (for example, an alpha particle's track is broad and straight, while that of an electron is thinner and shows more evidence of deflection). The tracks are photographed and analyzed. When a vertical magnetic field is applied, positively and negatively charged particles will curve in opposite directions. There are different types of the cloud

[3]The cloud chamber was invented by a Scottish physicist Charles Thomas Rees Wilson (1869–1959)

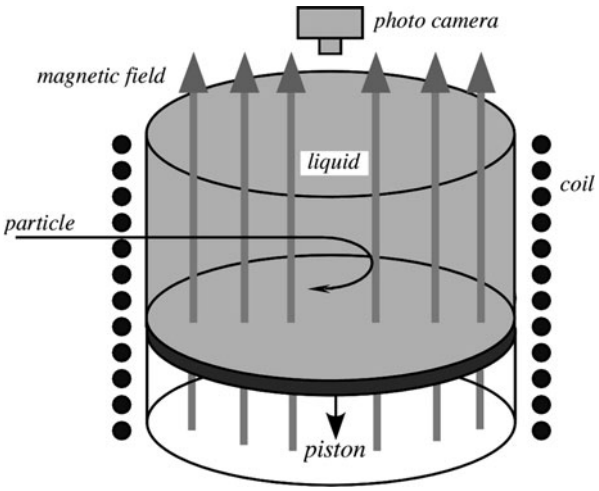

Fig. 15.7 Bubble chamber

chambers. The expansion cloud chambers use a vacuum pump to briefly produce the right conditions for trails to form, while the diffusion type uses solid CO_2 (dry ice) to cool the bottom of the chamber and produce a temperature gradient in which trails can be seen continuously.

A bubble chamber[4] is similar to the cloud chamber except that a liquid is used instead of vapor. Interestingly, a glass of champagne or beer is a kind of a bubble chamber where tiny bubble formations are triggered by the ionizing radiation coming from the environment and outer Space. For the physical experiments, a bubble chamber is filled with a more prosaic and much colder liquid, such as liquid hydrogen. It is used for detecting electrically charged particles moving through it.

The bubble chamber is normally made by filling a large cylinder (Fig. 15.7) with liquid hydrogen heated to just below its boiling point. As particles enter the chamber, a piston suddenly decreases its pressure, and the liquid enters into a superheated metastable phase. Charged particles create an ionization track, around which the liquid vaporizes, forming microscopic bubbles. Bubble density around a track is proportional to a particle's energy loss.

Bubbles grow in size as the chamber expands, until they are large enough to be seen or photographed. Several cameras are mounted around it, allowing a three-dimensional image of an event to be captured. Bubble chambers with resolutions down to a few μm have been operated. A constant magnetic field is formed around the chamber by an electromagnet that causes charged particles to travel in helical paths whose radii are determined by their charge-to-mass ratios. Although bubble chambers were very successful in the past, they are of only limited use in current

[4]The bubble chamber was invented in 1952 by Donald A. Glaser, for which in 1960 he was awarded the Nobel Prize in Physics.

very-high-energy experiments for a variety of reasons, some of them are the problem with the superheated phase that must be ready at the precise moment of collision, which complicates the detection of short-lived particles. Also, the bubble chambers are neither large nor massive enough to analyze high-energy collisions, where all products should be contained inside the detector.

References

1. Evans RD (1955) The atomic nucleus. McGraw-Hill, New York
2. Knoll GF (1999) Radiation detection and measurement. 3rd Ed., Wiley, New York

Chapter 16
Temperature Sensors

When a scientist thinks of something, he asks, – "Why?"
When an engineer thinks of something, he asks, – "Why not?"

From prehistoric times people were aware of heat and trying to assess its intensity by measuring temperature. Perhaps the simplest and certainly the most widely used phenomenon for temperature sensing is thermal expansion. This forms the basis of the liquid-in-glass thermometers. For the electrical transduction, different methods of sensing are employed. Some of them are the resistive, thermoelectric, semiconductive, optical, acoustic, and piezoelectric detectors.

16.1 Coupling with Object

Taking a temperature essentially requires the transmission of a small portion of the object's thermal energy to the sensor whose function is to convert that energy into an electrical signal. When a contact sensor (probe) is placed inside or on the object, heat conduction takes place through the boundary between the object and the probe. The sensing element in the probe warms up or cools down, i.e., it exchanges heat with the object. The same happens when heat is transferred by means of radiation – thermal energy in the form of infrared light is exchanged between the sensor and the object. Any sensor, no matter how small, will disturb the measurement site and thus cause some error in temperature measurement. This applies to any method of sensing: conductive, convective, and radiative. Thus, it is an engineering task to minimize the error by an appropriate sensor design and a correct measurement technique of which the coupling between the sensor and object is most critical.

When a contact temperature sensor responds to heat, two basic methods of the signal processing can be employed: equilibrium and predictive. In the equilibrium method, a temperature measurement is complete when no significant thermal

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_16, © Springer Science+Business Media, LLC 2010

gradient exists between the measured surface and the sensing element inside the probe. In other words, a thermal equilibrium is reached between the sensor and object of measurement when energy exchange becomes negligible. In the predictive method, the equilibrium is not reached during the measurement time. The equilibrium level is anticipated via a computation through the rate of the sensor's temperature change. After the initial probe placement, reaching a thermal equilibrium between the object and sensor may be a slow process, especially if the contact area is dry. Hence, the process of temperature equalization may take significant time. For instance, a contact medical electronic thermometer may take temperature from a water bath within about 5 s (good thermal coupling), but it will be at least 3 min when temperature is measured axillary (under the armpit) by the same probe (poor thermal coupling).

Let us discuss what affects the accuracy of a temperature measurement. If a sensor is coupled not only to the object whose temperature is measured, but to some other items as well, an error is introduced. To be sure, a temperature sensor is *always* attached to something else besides the object of measurement. An example of another item is a connecting cable (Fig. 16.1a). The sensor is coupled to the object (with an adhesive, for example). After the moment of coupling or after the object temperature changes, the sensor's temperature at any moment of time is T_S, while the object has true temperature T_B. Almost never these two temperatures are really equal. The goal of the equilibrium measurement is to bring T_S as close to T_B as possible. In any practical system, one end of the cable is connected to the sensor while the other end is subjected to another temperature, for example, the ambient temperature T_0 that may be quite different from that of the object. The cable conducts both an electric signal and some heat from or to the sensor. Figure 16.1b shows a thermal circuit that includes the object, sensor, environment, and thermal resistances r_1 and r_2. Thermal resistances should be clearly understood. A thermal resistance represents the ability of a matter to conduct thermal energy and is inversely related to thermal conductivity, that is, $r = 1/\alpha$. If an object is warmer than the environment, heat flows in the direction indicated by an arrow.

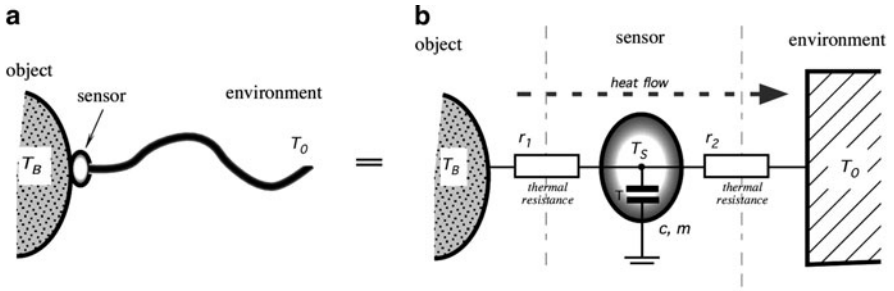

Fig. 16.1 Temperature sensor has thermal contacts with both the object and the connecting cable (**a**); equivalent thermal circuit (**b**)

The circuit of Fig. 16.1b resembles an electric circuit and indeed its properties can be evaluated by using the laws of electric circuits such as Kirchhoff's[1] and Ohm's laws. Note that a thermal capacitance is represented by a capacitor. Assuming that we wait sufficiently long and all temperatures are settled on some steady-state levels, and also assuming that the object and environment temperatures are stable and not affected by their interconnection by the sensor, for such a steady state, we may apply the law of conservation of energy. Consider that the thermal energy that flows from the object to the sensor is equal to the energy that outflows from the sensor to the environment. This allows us to write the balance equation:

$$\frac{T_B - T_S}{r_1} = \frac{T_B - T_0}{r_1 + r_2} \tag{16.1}$$

from which we derive the sensor's temperature as

$$T_S = T_B - (T_B - T_0)\frac{r_1}{r_2} = T_B - \Delta T\frac{r_1}{r_2}, \tag{16.2}$$

where ΔT is a thermal gradient between the object and the surroundings. Let us take a closer look at (16.2). We can draw several conclusions from it. The first is that the sensor temperature T_s is different from that of the object. The only exception is when the environment is at the same temperature as the object (a special case when $\Delta T = T_{Bc} - T_0 = 0$). The second conclusion is that T_s will closely approach T_B regardless of the temperature gradient ΔT when the ratio r_1/r_2 approaches zero. This means that for minimizing the measurement error one must improve a thermal coupling between the object and the sensor and decouple the sensor from the surroundings as much as practical. Often, it is not easy to do.

The best way to bring ΔT closer to zero is to imbed the sensor into the object as shown in Fig. 16.2b. A cavity is formed inside the object where the sensor is placed, preferably with a thermal grease, epoxy, or other method is used the thermally bond the sensor with the cavity walls. The cable near the sensor is formed as a loop and also placed inside the cavity. This allows equalizing temperatures of the sensor, cable and cavity. Since the sensor and portion of the cable are not exposed to the external temperature, the measurement becomes significantly more accurate.

Forming a cavity inside the object is not always possible that makes the surface measurement the only practical choice. This is not a desirable arrangement. Yet, there is a powerful technique to form a "virtual cavity" at the object's surface, forcing $\Delta T \to 0$. In a practical circuit, the prime sensor is provided with a thermal shied shown in Fig. 16.2b. This is a conceptual equivalent to a driven capacitive shield (for example, see Figs. 5.4b and 6.7). The thermal shield is fabricated of a metal having a good thermal conductivity (e.g., aluminum) and contains the imbedded heater (and/or cooler) and another temperature sensor called the shield

[1]Kirchhoff's law was originally developed not for the electrical circuits but for plumbing.

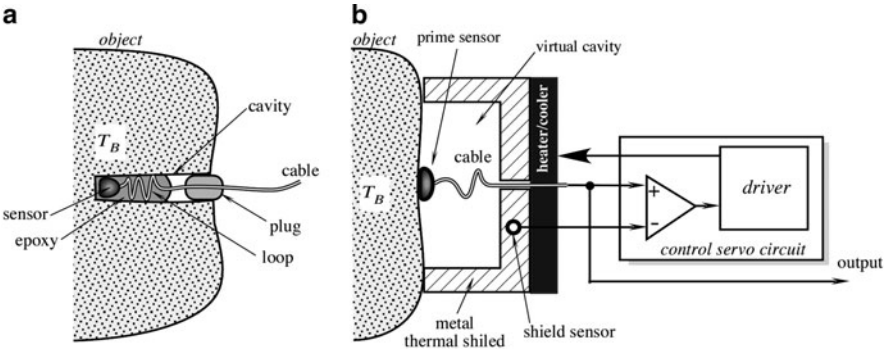

Fig. 16.2 Imbedded temperature sensor (**a**) and surface temperature sensor with an active driven thermal shield (**b**)

temperature sensor. Both sensors provide signals to the control servo-circuit that supplies power to the heater/cooler. The servo-circuit works to minimize the thermal gradient ΔT between the prime sensor and shield sensors. Preferably, the shield touches the object and thus protects the prime sensor from the environment. The object surface and the shield form a virtual thermostatic cavity around the measurement site. When the thermal gradient between both sensors approaches zero, the prime sensor becomes nearly totally thermally decoupled from the environment. It is very important to thermally decouple the prime sensor from the thermal shield; otherwise, the circuit may become unstable. This method was implemented in a medical body core thermometer. The prime sensor was placed inside the ear canal touching the skin, while the thermal shield with a heater was attached to the ear helix [1]. This sensor allowed accurate continuous noninvasive monitoring of temperature inside the patient's head without penetrating the scull.

Above, we evaluated the static condition of the sensor, now let us consider a dynamic case when temperatures change with time. This occurs when either the object or the surrounding temperatures change or the sensor was recently attached to the object and its temperature is not yet stabilized.

Initially, let us consider an ideal case that requires making two assumptions: (1) A thermal resistance between the sensor and the environment is infinitely large ($r_2 \to \infty$) and (2) the object's temperature does not change after the sensor is attached. In other words, the object is considered being much larger than the sensor and acting as an "infinite" heat source/sink. In other words, it is having an infinitely large thermal capacity and infinitely large thermal conductivity. At a starting time $t = 0$, the temperature sensor having the initial temperature T_1 comes in contact with the object having temperature T_B. After that according to Newton's law of cooling the incremental amount of transferred heat to the sensor is proportional to a temperature gradient between the instant sensor temperature T_s at a particular moment and temperature of the object T_B:

$$dQ = \alpha_1 (T_B - T_S) dt, \tag{16.3}$$

where $\alpha_1 = 1/r_1$ is the thermal conductivity of the sensor–object boundary. Note that T_s is changing. If the sensor has an average specific heat c and mass m, the heat absorbed by the sensor is

$$dQ = mcdT. \tag{16.4}$$

Equations (16.3) and (16.4) are equal and yield the first-order differential equation

$$\alpha_1(T_B - T_s)dt = mcdT. \tag{16.5}$$

We denote thermal time constant τ_T as

$$\tau_T = \frac{mc}{\alpha_1} = mcr_1, \tag{16.6}$$

then the differential equation takes form

$$\frac{dT}{T_B - T_s} = \frac{dt}{\tau_T}. \tag{16.7}$$

This equation has a solution

$$T_s = T_B - \Delta T e^{-\frac{t}{\tau_T}}. \tag{16.8}$$

The time transient of the sensor's temperature that corresponds to the above solution is shown in Fig. 16.3. One time constant τ_T is equal to the time required for temperature T to reach about 63.2% of the initial gradient $\Delta T = T_b - T_1$. The smaller the time constant the faster the sensor responds to a change in temperature. The time constant can be minimized by reducing the size of the sensor (smaller m) and by improving its coupling with the object (smaller r_1).

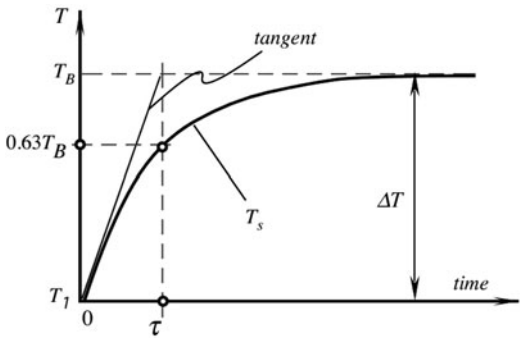

Fig. 16.3 Temperature changes of a sensor (the sensor is ideally coupled to an ideal object)

If we wait for a long time ($t \to \infty$) after attaching the sensor, (16.8) states then temperature of the sensor approaches temperature of the object: $T_s = T_B$ and the sensor reading can be used to compute the object's temperature. Theoretically, it takes infinite time to reach a perfect equilibrium between T_s and T_B – hardly anyone can wait that long! Fortunately, since only a finite accuracy is usually required, for most practical cases a quasi-equilibrium state may be considered after 5–10 time constants. For instance, after the waiting time $t = 5\tau$, the sensor's temperature will differ from that of the object by 0.7% of the initial gradient, while after 10 time constants it will be within 0.005%.

Now, we will study a more realistic case. Let us remove the first of the above assumptions and consider that a thermal coupling with the environment is not extremely large, that is, $r_2 \neq \infty$. Then, the thermal time constant should be determined from

$$\tau_T = \frac{mc}{\alpha_1 + \alpha_2} = mc\,\frac{r_1}{1 + \frac{r_1}{r_2}} \tag{16.9}$$

and the sensor's response is shown in Fig. 16.3a. Note that now the sensor temperature never reaches exactly that of the object, no matter how long we wait. Hence, even at the equilibrium state, there will be a remaining thermal gradient δT, which is the measurement error due to a poor decoupling from the environment.

Now, we will remove the second assumption; we consider the object not an ideal heat source or sink. This means that the object is not dramatically larger than the sensor or its thermal conductivity is not very large. As a result of this "imperfection," the sensor will disturb, at least temporarily, the measurement site after its attachment. Figure 16.4 shows that upon the sensor attachment, the object's temperature at the point of contact deflects and then will gradually return to some steady-state level. This causes a deviation of the sensor's temperature profile from the ideal exponential function and a concept of a thermal time constant (τ_T) will be no longer applicable. In practice, this deviation becomes significant if one wishes to employ a predictive algorithm as mentioned above, or a fast temperature tracking is

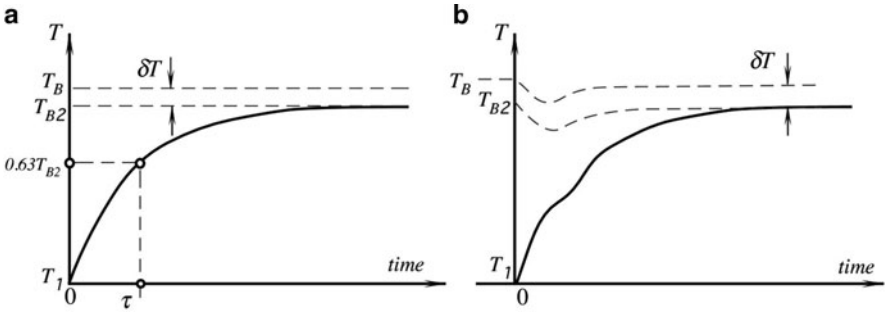

Fig. 16.4 Temperature changes of a sensor that is coupled to the environment (**a**) and when the object has a limited thermal conductivity (**b**)

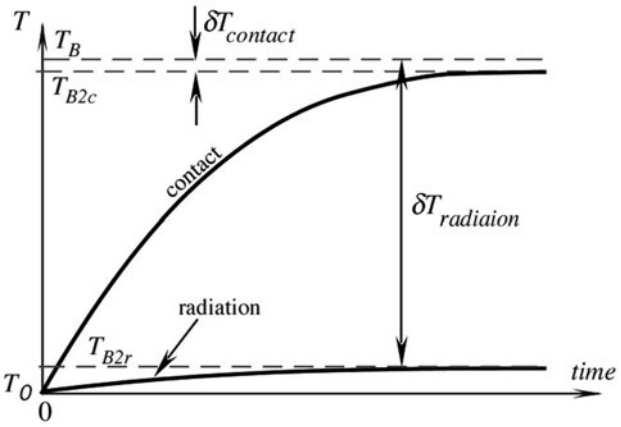

Fig. 16.5 Difference in thermal responses between contact and noncontact (IR radiation) temperature sensors

required. Predictive algorithms require knowledge of the sensor dynamic response function. The algorithm will be resulting in large errors if the real function differs from the preprogrammed functional. Again, the best known solutions for this problem are to use either an imbedded sensor or a dynamic thermal shield [1] that are shown in Fig. 16.2.

Note that the above discussion is equally applicable to all temperature sensors – either contact or noncontact infrared. The only difference is that for a contact sensor, typically $r_1 \ll r_2$, while for a noncontact IR sensor $r_1 \gg r_2$. In other words, a typical contact sensor is much better thermally coupled with the object than with the environment, while the IR sensor is very poorly thermally coupled with the object and much stronger coupled with its immediate surroundings (environment).[2] This difference is illustrated in Fig. 16.5.

A typical contact temperature sensor consists of the following components (Fig. 16.6a):

1. A sensing element – a device whose electrical properties vary somewhat in response to changes in temperature. A good sensing element should have low specific heat, very small mass, high thermal conductivity, and strong and predictable sensitivity to temperature.
2. Terminals are the conductive pads or wires that interface between the sensing element and the external electronic circuit. The terminals should have the lowest possible thermal conductivity and low electrical resistance (platinum is often the best compromise, yet expensive). Also, the terminals often are used to support the sensing element so they should have a reasonable mechanical strength and stability.

[2]This is the reason why an IR thermometer requires an ambient sensor that is thermally connected to the IR sensor.

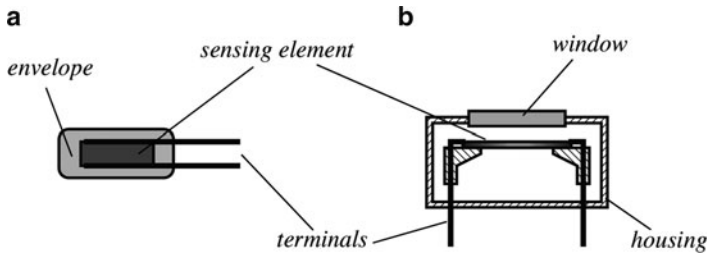

Fig. 16.6 General structures of temperature sensors: Contact sensor (**a**) and noncontact thermal IR radiation sensor (**b**)

3. A protective envelope is either housing or coating, which physically separates the sensing element from the environment. A good envelope must have low thermal resistance (high thermal conductivity) and high electrical isolation properties. It must be environmentally stable and impermeable to moisture and other compounds that may spuriously affect the sensing element.

A noncontact temperature sensor (Fig. 16.6b) is an optical thermal radiation sensor whose designs are covered in detail in Chap. 13. Like a contact sensor, it also contains a sensing element that is responsive to its own temperature. The difference is in the way of a heat transfer from an object to the element: in a contact sensor it is through thermal conduction by way of a physical contact, while in a noncontact sensor it is through thermal radiation (optically).

To improve a response time of a thermal radiation sensor, thickness of the sensing element is minimized, while for a better sensitivity its surface area is maximized. In addition to a sensing element, the noncontact thermal sensor may have an optical window and a built-in interface circuit. The interior of the sensor's housing is usually filled with dry air or nitrogen.

All temperature sensors can be divided into two classes: the absolute sensors and the relative sensors. An absolute temperature sensor measures temperature that is referenced to the absolute zero or any other point on the absolute temperature scale, for example 0°C (273.15°K), 25°C or a selected calibration temperature. Examples of the absolute sensors are thermistors and resistance temperature detectors (RTDs). A relative sensor measures a temperature difference between two objects where one object is called the reference. An example of a relative sensor is a thermocouple.

16.2 Temperature Reference Points

For calibration of any temperature sensor, a precision reference is required. Typically, a reference sensor is some kind of a very stable reference probe that, in turn, must be calibrated to even more stable and predictable reference. In science and

Table 16.1 Temperature reference points

Point description	°C
Triple point[a] of hydrogen	−259.34
Boiling point of normal hydrogen	−252.753
Triple point of oxygen	−218.789
Boiling point of nitrogen	−195.806
Triple point of argon	−189.352
Boiling point of oxygen	−182.962
Sublimation point of carbon dioxide	−78.476
Freezing point of mercury	−38.836
Triple point of water	0.01
Freezing point of water (water–ice mixture)	0.00
Boiling point of water	100.00
Triple point of benzoic acid	122.37
Freezing point of indium	156.634
Freezing point of tin	231.968
Freezing point of bismuth	271.442
Freezing point of cadmium	321.108
Freezing point of lead	327.502
Freezing point of zinc	419.58
Freezing point of antimony	630.755
Freezing point of aluminum	660.46
Freezing point of silver	961.93
Freezing point of gold	1,064.43
Freezing point of copper	1,084.88
Freezing point of nickel	1,455
Freezing point of palladium	1,554
Freezing point of platinum	1,769

[a]Triple point is equilibrium between the solid, liquid, and vapor phases.

industry, such references are certain chemical compounds whose temperature behavior at selected equilibrium states is governed by the fundamental laws of nature. During calibration, the reference thermometer is placed at a controlled pressure inside the material being at a specific state (Table 16.1) and the sensor response is measured. Then, it is moved to the next material and calibrated again. After being calibrated at several temperature points, the sensor may be served as a secondary standard reference.

The calibration scale depends on the selected standard. According to the International Temperature Scale (ITS-90),[3] precision temperature instruments should be calibrated at reproducible equilibrium states of some materials. This scale designated kelvin temperatures by symbol T_{90} and the Celsius scale by t_{90}.

[3]The International Temperature Scale of 1990 was adopted by the International Committee of Weights and Measures at its meeting in 1989. This scale supersedes the International Practical Temperature Scale of 1968 (amended edition of 1975) and the 1976 Provisional Temperature Scale.

16.3 Thermoresistive Sensors[4]

Sir Humphry Davy had noted as early as 1821 that electrical resistances of various metals depend on temperature [2]. Sir William Siemens, in 1871, first outlined the use of a platinum resistance thermometer. In 1887, Hugh Callendar published a paper [3] where he described how to practically use platinum temperature sensors. The advantages of thermoresistive sensors are in simplicity of the interface circuits, sensitivity, and long-term stability. All such sensors can be divided into three groups: RTDs, pn-junction detectors, and thermistors. They belong to class of the *absolute* temperature sensors, that is, they can measure temperatures that are referenced to an absolute temperature scale.

16.3.1 Resistance Temperature Detectors

This term is usually pertinent to metal sensors fabricated in the form of either a wire or a thin film. Nowadays, this class also covers some semiconductor materials with a pronounced sensitivity to temperature (e.g., germanium). Temperature dependence of resistivities of all metals and most alloys gives an opportunity to use them for temperature sensing (Table A.7). While virtually all metals can be employed for sensing, platinum is used almost exclusively because of its predictable response, long-term stability, and durability. Tungsten RTDs are usually applicable for temperatures over 600°C. All RTDs have positive temperature coefficients (PTCs). Several types of them are available from various manufacturers:

1. Thin-film RTDs are often fabricated of thin platinum or its alloys and deposited on a suitable substrate such as a micromachined silicon membrane. The RTD is often made in a serpentine shape to ensure a sufficiently large length/width ratio.
2. Wire-wound RTDs, where the platinum winding is partially supported by a high-temperature glass adhesive inside a ceramic tube. This construction provides a detector with the most stability for industrial and scientific applications.

Equation (3.58) gives a best fit second-order approximation for platinum. In industry, it is customary to use separate approximations for the cold and hot temperatures. Callendar–van Dusen approximations represent platinum transfer functions:

For the range from −200°C to 0°C

$$R_t = R_0\left[1 + At + Bt^2 + Ct^3(t - 100)\right]. \tag{16.10}$$

[4]Also, see Sect. 3.11.

For the range from 0°C to 630°C is becomes identical to (3.58)

$$R_t = R_0\left[1 + At + Bt^2\right]. \tag{16.11}$$

The constants A, B, and C are determined by the properties of platinum used in the construction of the sensor. Alternatively, the Callendar–van Dusen approximation can be written as

$$R_t = R_0\left\{1 + \alpha\left[1 - \delta\left(\frac{t}{100}\right)\left(\frac{t}{100} - 1\right) - \beta\left(\frac{t}{100}\right)^3\left(\frac{t}{100} - 1\right)\right]\right\}, \tag{16.12}$$

where t is the temperature in °C and the coefficients are related to A, B, and C as

$$A = \alpha\left(1 + \frac{\delta}{100}\right), B = -\alpha\delta \times 10^{-4}, C = -\alpha\beta \times 10^{-8}. \tag{16.13}$$

The value of δ is obtained by calibration at a high temperature, for example, at the freezing point of zinc (419.58°C), and β is obtained at the calibration at a negative temperature.

To conform with the ITS-90, the Callendar–van Dusen approximation must be corrected. The correction is rather complex and the user should refer for details to the ITS-90. In different countries, some national specifications are applicable to RTDs. For instance, in Europe, these are BS 1904:1984; DIN 43760-1980; IEC 751:1983. In Japan, it is JISC1604-1981. In the United States, different companies have developed their own standards for α values. For example, SAMA Standard RC21-4-1966 specifies $\alpha = 0.003923°C^{-1}$, while in Europe DIN standard specifies $\alpha = 0.003850°C^{-1}$, and the British Aircraft industry standard is $\alpha = 0.003900°C^{-1}$.

Usually, RTDs are calibrated at standard points, which can be reproduced in a laboratory with high accuracy (Table 16.1). Calibrating at these points allows for precise determination of approximation constants α and δ.

Typical tolerances for the wire-wound RTDs is ±10 mΩ, which corresponds to about ±0.025°C. Giving high requirements to accuracy, packaging isolation of the device should be seriously considered. This is especially true at higher temperatures where the resistance of isolators may drop significantly. For instance, a 10 MΩ shunt resistor at 550°C results in resistive error of about 3 mΩ, which corresponds to the temperature error of −0.0075°C.

16.3.2 Silicon *Resistive PTC Sensors*

Conductive properties of bulk silicon have been successfully implemented for fabrication of temperature sensors with PTC characteristics. Nowadays, silicon resistive sensors are often incorporated into the micromachined structures for

temperature compensation or direct temperature measurement. There are also the discrete silicon sensors, for example the so-called KTY temperature detectors that originally were manufactured by Philips. Nowadays, these sensors are also produced by other manufacturers, for example (www.amwei.com) and (www.rtie.com). The Si PTC sensors have reasonably good linearity and high, long-term stability (typically ±0.05 K per year). The PTC makes them inherently safe for operation in heating systems – a moderate overheating (below 200°C) results in RTD's resistance increase and a self-protection. Silicon RTD belong to the class of absolute temperature sensors, that is, they can measure temperature that is referenced to an absolute temperature scale.

Pure silicon, either polysilicon or single crystal silicon, intrinsically has a negative temperature coefficient (NTC) of resistance (Fig. 18.1b). However, when it is doped with an n-type impurity, in a certain temperature range, its temperature coefficient becomes positive (Fig. 16.7). This is the result of the fall in charge carrier mobility at lower temperatures. At higher temperatures, the number n of free charge carriers increases due to the number n_i of spontaneously generated charge carriers and the intrinsic semiconductor properties of silicon predominate. Thus, at temperatures below 200°C, resistivity ρ has a PTC, while over 200°C it becomes negative. The basic KTY sensor consists of an n-type silicon cell having approximate dimensions of 500 × 500 × 240 µm, metallized on one side and having contact areas on the other side. This produces an effect of resistance "spreading," which causes a conical current distribution through the crystal, significantly reducing the sensor's dependence on manufacturing tolerances. A KTY sensor may be somewhat sensitive to a current direction, especially at larger

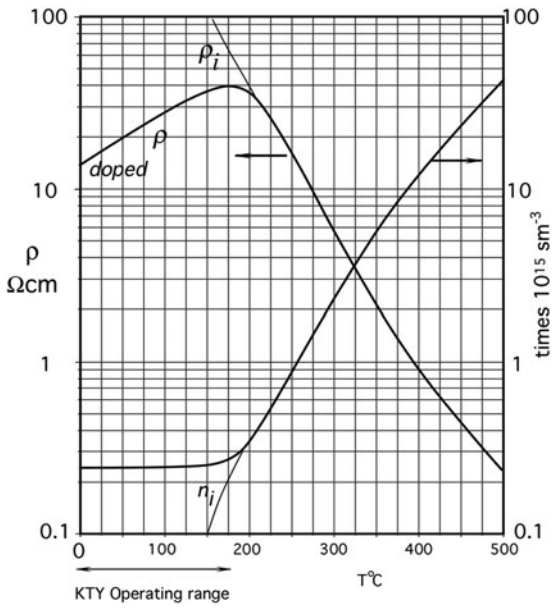

Fig. 16.7 Resistivity and number of free charge carriers for n-doped silicon

currents and higher temperatures. To alleviate this problem, a serially opposite design is employed where two of the sensors are connected with opposite polarities to form a dual sensor. These sensors are especially useful for the automotive applications.

A typical sensitivity of a PTC silicon sensor is on the order of 0.7%/°C, that is, its resistance changes by 0.7% per every degree C. As for any other sensor with a mild nonlinearity, the KTY sensor transfer function may be approximated by a second-order polynomial

$$R_T = R_0\left[1 + A(T - T_0) + B(T - T_0)^2\right], \tag{16.14}$$

where R_0 and T_0 are the resistance (Ω) and temperature (K) at a reference point. For instance, for the KTY-81 sensors operating in the range from −55 to +150°C, the coefficients are: $A = 0.007874\ \mathrm{K}^{-1}$ and $B = 1.874 \times 10^{-5}\ \mathrm{K}^{-2}$. A typical transfer function of the sensor is shown in Fig. 16.8.

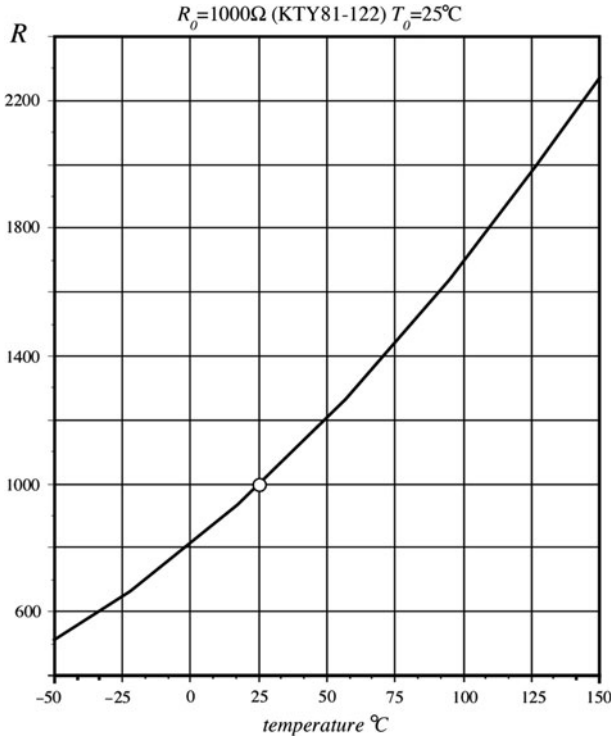

Fig. 16.8 Transfer function of a KTY silicon temperature sensor

16.3.3 Thermistors

The term *thermistor* is a contraction of words *thermal* and *resistor*. The name is usually applied to metal-oxide sensors fabricated in forms of droplets, bars, cylinders, rectangular flakes, and thick films. Thermistors can be also fabricated of silicon and germanium. A thermistor belongs to class of the *absolute* temperature sensors, that is, it can measure temperature that is referenced to an absolute temperature scale. All thermistors are divided into two groups: NTC and PTC. Only the NTC thermistors are useful for precision temperature measurements.

16.3.3.1 NTC Thermistors

A conventional metal-oxide thermistor has a NTC, that is, its resistance decreases with the increase in temperature. The NTC thermistor's resistance, as of any resistor, is determined by its physical dimensions and the material resistivity. The relationship between the resistance and temperature is highly nonlinear (Fig. 3.18).

Whenever a high accuracy is required, or the operating temperature range is wide, thermistor characteristics should not be taken directly from a manufacturer's data sheet. Typical tolerances of the nominal resistance (at 25°C) for the mass-produced thermistors are rather wide: ±10% is quite common, however, for a higher price a 1% or even better thermistors are readily available. Unless it was adjusted at a factory to a better tolerance, to reach a high accuracy, each low-tolerance thermistor needs to be individually calibrated over the entire operating temperature range. Manufacturers can trim a thermistor by grinding its body to a required dimension that directly controls the nominal value of resistance at a set temperature. This, however, increases cost. An alternative approach for an end user is to individually calibrate the thermistors. Calibration means that a thermistor has to be subjected to a precisely known temperature (a stirred water bath is often employed[5]) and its resistance is measured. This is repeated at several temperatures if a multipoint calibration is needed (see Sect. 2.2). Naturally, a thermistor calibration is as good as the accuracy of a reference thermometer used during the calibration. To measure resistance of a thermistor, it is attached to a measurement circuit that passes through it an electric current. Depending on the required accuracy and the production cost restrictions, a thermistor calibration can be based on use of one of several known approximations (models) of its temperature response.

When a thermistor is used as a temperature sensor, we assume that all its characteristics are based on the so-called zero-power resistance, meaning that electric current passing through a thermistor does not result in any noticeable temperature increase (self-heating) that may affect accuracy of measurement.

[5]Actually, water is not used. Mineral oil or Fluorinert® electronic fluid is more of practical liquid.

A static temperature increase in a thermistor due to self-heating is governed by the following equation:

$$\Delta T_H = r\frac{N^2V^2}{S}, \tag{16.15}$$

where r is a thermal resistance to surroundings, V is the applied dc voltage during the resistance measurement, S is the resistance of a thermistor at a measured temperature, and N is a duty cycle of measurement (for example, $N = 0.1$ means that constant voltage is applied to a thermistor only during 10% of the time). For a dc measurement, $N = 1$.

As follows from (6.15), a zero-power can be approached by selecting high-resistance thermistors, increasing coupling to the object of measurement (reducing r), and measuring its resistance at low voltages applied during short time intervals. At the end of this chapter, we will show the effects of self-heating on the thermistor response, but for now, we assume that a self-heating results in a negligibly small error.

To use a thermistor in the actual device, its transfer function (temperature dependence of a resistance) must be accurately established. Since that function is highly nonlinear (Fig. 16.15) and generally specific for each particular sensor, an analytical equation connecting the resistance and temperature is highly desirable. Several mathematical models of a thermistor transfer function have been proposed. It should be remembered, however, that any model is only an approximation, and generally the simpler the model, the lesser the accuracy should be expected. On the other hand, at a more complex model, calibration and the use of a thermistor become more difficult. All present models are based on the experimentally established fact that logarithm of a thermistor's resistance S relates to its absolute temperature T by a polynomial equation

$$\ln S = A_0 + \frac{A_1}{T} + \frac{A_2}{T^2} + \frac{A_3}{T^3}. \tag{16.16}$$

From this basic equation, three computational models have been proposed.

16.3.3.2 Simple Model

Over a relatively narrow temperature range and assuming that some accuracy may be lost, we can eliminate two last terms in (16.16) and arrive at [4]:

$$\ln S \cong A + \frac{\beta_m}{T}, \tag{16.17}$$

where A is a dimensionless constant, and β_m is another constant called the material characteristic temperature (in kelvin). If a thermistor's resistance S_0 at a calibrating temperature T_0 is known, then the resistance-temperature relationship is expressed as

$$S = S_0 \mathrm{e}^{\beta_m \left(\frac{1}{T} - \frac{1}{T_0}\right)}. \tag{16.18}$$

Equation (16.18) is the most popular and widely used thermistor model. An obvious advantage of this model is a need to calibrate a thermistor at only one point (S_0 at T_0). However, this assumes that the value of β_m is known beforehand; otherwise, a two-point calibration is required to find the value of β_m:

$$\beta_m = \frac{\ln \frac{S_1}{S_0}}{\left(\frac{1}{T_1} - \frac{1}{T_0}\right)}, \tag{16.19}$$

where T_0 and S_0, T_1 and S_1 are two pairs of the corresponding temperatures and resistances at two calibrating points on the curve described by (16.18). The value of β_m is considered temperature independent, but may vary from part to part due to the manufacturing tolerances, which typically are within $\pm 1\%$.

When the thermistor is used as a sensor, its resistance S is measured. From that resistance, temperature can be computed as

$$T = \left(\frac{1}{T_0} + \frac{\ln \frac{S}{S_0}}{\beta_m}\right)^{-1}. \tag{16.20}$$

Error in temperature computation from the approximation provided by (16.20) is small near the calibrating temperature T_0, but it increases noticeably while moving away from that point (Fig. 16.10).

Beta specifies a thermistor curvature, but it does not directly describe its sensitivity (NTC), which is the NTC α. The coefficient can be found by differentiating and normalizing (16.18)

$$\alpha_r = \frac{1}{S}\frac{dS}{dT} = -\frac{\beta}{T^2}. \tag{16.21}$$

It follows from (16.21) that the sensitivity depends on both: beta and temperature. A thermistor is much more sensitive at lower temperatures while its sensitivity drops fast with a temperature increase. Equation (16.21) shows what fraction of the resistance S changes per one degree of temperature. In the ceramic NTC thermistors, the sensitivity α varies over the temperature range from -2% (at the warmer side of the scale) to $-8\%/^{\circ}\mathrm{C}$ (at the cooler side of the scale), which implies that an NTC thermistor albeit a nonlinear sensor is a very sensitive device, roughly an order of magnitude more temperature sensitive than any RTD. This is especially

important for applications where a high-output signal over a relatively narrow temperature range is desirable. An example is a medical electronic thermometer.

16.3.3.3 Fraden Model

In 1998, the author of this book proposed a further improvement of the simple model [5]. It is based on the experimental fact that characteristic temperature β is not a constant but rather a function of temperature (Fig. 16.9). Depending on the manufacturer and type of a thermistor, the function may have either positive slope, as shown in the picture, or negative one. Ideally, β should not change with temperature, but that is just a special case that can be seen only from the best manufacturers who tightly control composition of the ceramic material. In such rare cases, the simple model provides an accurate basis for temperature computation. But for a relatively inexpensive sensor, the Fraden model should be considered.

It follows from (16.16) and (16.17) that the thermistor material characteristic temperature β can be approximated as

$$\beta = A_1 + BT + \frac{A_2}{T} + \frac{A_3}{T^2}, \tag{16.22}$$

where A and B are constants. The evaluation of this equation shows that the third and fourth summands are very small when compared with the first two and for most

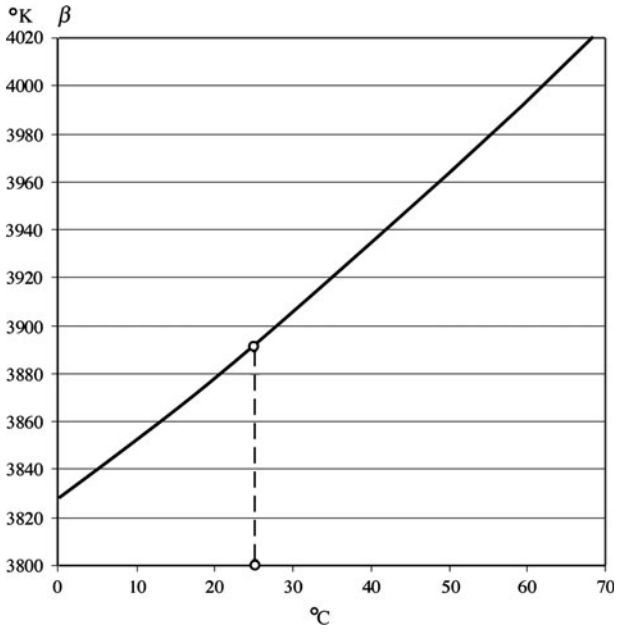

Fig. 16.9 Value of β changes with temperature

practical cases can be removed. After elimination of two last terms, a model for the material constant β can be represented as linear function of temperature:

$$\beta = A_1 + BT. \tag{16.23}$$

Considering β as linear function of temperature, the simple model can be refined to improve its fidelity. Since β is no longer a constant, its linear function can be defined through at least one fixed point at some selected temperature T_b and a slope γ. Then, (16.23) can be written in form

$$\beta = \beta_b[1 + \gamma(T - T_b)], \tag{16.24}$$

where β_b is attributed to temperature T_b. A dimensionless coefficient γ has a meaning of a normalized change (a slope) in β:

$$\gamma = \left(\frac{\beta_x}{\beta_y} - 1\right)\frac{1}{T_c - T_a}, \tag{16.25}$$

where β_x and β_y are two material characteristic temperatures at two T_a and T_c characterizing temperatures.[6] To determine γ, three characterizing temperature points are required (T_a, T_b, and T_c); however, the value of γ does not need to be characterized for each individual thermistor. The value of γ depends on the thermistor material and the manufacturing process; so, it may be considered more or less constant for a production lot of a particular type of a thermistor. Thus, it is usually sufficient to find γ for a lot or type of a thermistor rather than for each individual sensor.

By substituting (16.23) into (16.16), we arrive at a model of a thermistor:

$$\ln S \cong A + \frac{\beta_m[1 - \gamma(T_b - T)]}{T}. \tag{16.26}$$

Solving (16.26) for resistance S yields the equation representing the thermistor's resistance as function of its temperature:

$$S = S_0 \mathrm{e}^{\beta_m[1+\gamma(T-T_0)]\left(\frac{1}{T}-\frac{1}{T_0}\right)}, \tag{16.27}$$

where S_0 is the resistance at calibrating temperature T_0, and β_m is the characteristic temperature defined at two calibrating temperatures T_0 and T_1 (16.19). This is similar to a simple model of (16.18) with an introduction of an additional constant γ. Even though this model requires three points to define γ for a production lot, each

[6]Note that β and T are in kelvin. When temperature is indicated as t, the scale is in celsius.

individual thermistor needs to be calibrated at two points. This makes the Fraden model quite attractive for low-cost, high-volume applications that at the same time require higher accuracy. Note that the calibrating temperatures T_0 and T_1 preferably should be selected closer to the ends of the operating range, and for the characterization, temperature T_B should be selected near the middle of the operating range. See Table 16.2 for practical equations to use this model. Errors in temperature computation care shown in Fig. 16.10.

16.3.3.4 Steinhart–Hart Model

Steinhart and Hart in 1968 proposed a model for the oceanographic range from −3 to 30°C [6], which in fact is useful for a much broader range. The model is based on (16.16) from which temperature can be calculated as

$$T = \left[\alpha_0 + \alpha_1 \ln S + \alpha_2 (\ln S)^2 + \alpha_3 (\ln S)^3\right]^{-1}. \tag{16.28}$$

Steinhart and Hart showed that the square term can be dropped without any noticeable loss in accuracy so the final equation becomes

$$T = \left[b_0 + b_1 \ln S + b_3 (\ln S)^3\right]^{-1}. \tag{16.29}$$

A correct use of the above equation assures accuracy in a millidegree range from 0 to 70°C [7]. To find coefficients b for the above equation, a system of three equations should be solved after a thermistor is calibrated at three temperatures (Table 16.2). Thanks to a very close approximation, the Steinhart and Hart model became an industry standard for calibrating precision thermistors. Some manufacturers prefer to use the complete equation (16.28), while the others find the simplified version (16.29) more practical. Extensive investigation of the Steinhart–Hart accuracy has demonstrated that even over a broader temperature range, the approximation error does not exceed the measurement uncertainty of a couple of millidegrees [8]. Nevertheless, a practical implementation of the approximation for the mass produced instruments is significantly limited by the need to calibrate each sensor at three or four temperature points.

A practical selection of the appropriate model depends on the required accuracy and cost constrains. A cost is affected by the number of points at which the sensor must be calibrated. A calibration is time-consuming and thus expensive. A complexity of mathematical computations is not a big deal thanks to a computational power of modern microprocessors. When accuracy demand is not high, or cost is of a prime concern, or the application temperature range is narrow (typically ±5 to 10°C from the calibrating temperature), the simple model is sufficient. The Fraden model is preferred when low cost and higher accuracy is a must. The Steinhart–Hart

Table 16.2 Practical use of three NTC thermistor models

	Simple model	Fraden model	Steinhart–Hart model
Maximum error from 0°C to 70°C	±0.7°C	±0.03°C	±0.003°C
Number of characterizing temperatures	2	3	0
Number of calibrating temperatures	2	2	3
Resistance–temperature dependence	$S = S_0 e^{\beta_m\left(\frac{1}{T}-\frac{1}{T_0}\right)}$	$S = S_0 e^{\beta_0[1+\gamma(T-T_0)]\left(\frac{1}{T}-\frac{1}{T_0}\right)}$	$S = e^{\left(A_0+\frac{A_1}{T}+\frac{A_2}{T^2}+\frac{A_3}{T^3}\right)}$
Characterizing a production lot or type of the thermistors			
Characterizing points	No characterization required for a two-point calibration	S_a at T_a, S_b at T_b, and S_c at T_c for a temperature range from T_a to T_c, where T_b is in the middle of the range	No characterization required
Characterizing factors		$\gamma = \left(\frac{\beta_x}{\beta_y}-1\right)\frac{1}{T_c-T_a}$, where $\beta_x = \frac{\ln\frac{S_c}{S_b}}{\left(\frac{1}{T_c}-\frac{1}{T_b}\right)}$, $\beta_y = \frac{\ln\frac{S_a}{S_b}}{\left(\frac{1}{T_a}-\frac{1}{T_b}\right)}$	
Calibrating an individual thermistor			
Calibrating Points	S_0 at T_0 and S_1 at T_1	S_0 at T_0 and S_1 at T_1	S_1 at T_1, S_2 at T_2, and S_3 at T_3
Analytic computation of temperature T (in kelvin) from resistance S			
Insert resistance S, the characterizing factors and the calibrating factors into the equation	$T = \left(\frac{1}{T_0}+\frac{ln\frac{S}{S_0}}{\beta_m}\right)^{-1}$ where $\beta_m = \frac{\ln\frac{S_1}{S_0}}{\left(\frac{1}{T_1}-\frac{1}{T_0}\right)}$	$T = \left(\frac{1}{T_0}+\frac{\ln\frac{S}{S_0}}{\beta_m[1-\gamma(T_1-T_r)]}\right)^{-1}$ where $T_r = \left(\frac{1}{T_0}+\frac{\ln\frac{S}{S_0}}{\beta_m}\right)^{-1}$ and $\beta_m = \frac{\ln\frac{S_1}{S_0}}{\left(\frac{1}{T_1}-\frac{1}{T_0}\right)}$	$T = \left[A + B\ln S + C(\ln S)^3\right]^{-1}$ where $C = \left(G-\frac{ZH}{F}\right)\left[\left(\ln S_1^3 - \ln S_2^3\right) - \frac{Z}{F}\left(\ln S_1^3 - \ln S_3^3\right)\right]^{-1}$ $B = Z^{-1}\left[G - C\left(\ln S_1^3 - \ln S_2^3\right)\right]$ $A = T_1^{-1} - C\ln S_1^3 - B\ln S_1$, $Z = \ln S_1 - \ln S_2, F = \ln S_1 - \ln S_3, H = T_1^{-1} - T_3^{-1}, G = T_1^{-1} - T_2^{-1}$

A thermistor type or lot should be characterized first to find the *characterizing* factor (Fraden model).
An individual thermistor is calibrated to determine the *calibrating* factors.
To compute a temperature T, measure the thermistor resistance S and calculate temperature with the use of characterizing and calibrating factors. All temperatures are in kelvin.

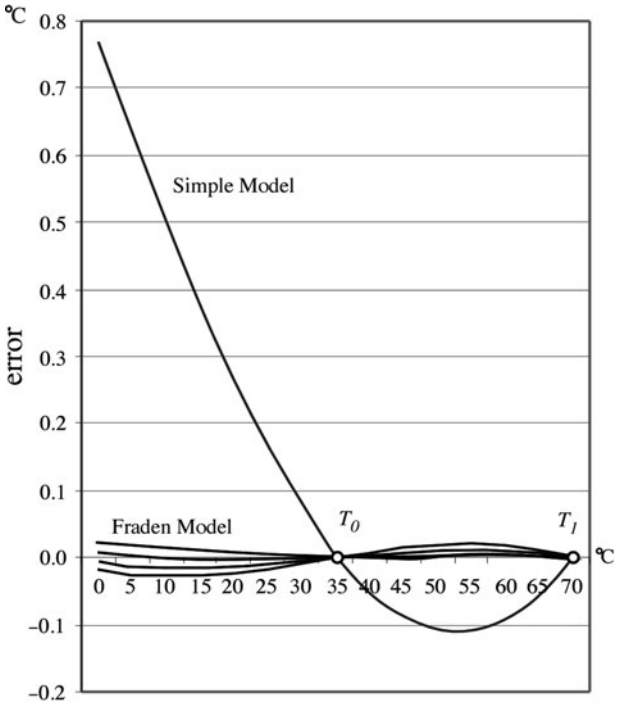

Fig. 16.10 Errors of a simple and Fraden models for a thermistor calibrated at two temperature points (t_0 and t_1) to determine β_m. Errors of a Steinhart–Hart model are too small to be shown on this scale

model should be used when the highest possible accuracy is required while the cost is not a major limiting factor.

To use a simple model, you need to know the values of β_m and the thermistor resistance S_0 at a calibrating temperature T_0. To use Fraden model, you need also know the value of γ, which is not unique for each thermistor but is unique for a lot or a type. For the Steinhart–Hart model, you need to know three resistances at three calibrating temperatures. Table 16.2 provides the equations for calibrating and computing temperatures from the thermistor resistances. For all three models, it requires a series of computations if the equations to be resolved directly. However, in most practical cases, these equations can be substituted by the look-up tables. To minimize the size of a look-up table, a piece-wise linear approximation can be employed (see Sections 2.1.5 and 2.3.1 in Chap. 2).

16.3.3.5 Fabrication of Ceramic NTC Thermistors

Generally, the NTC thermistors can be classified into three major groups depending upon the method by which they are fabricated. The first group consists of bead-type

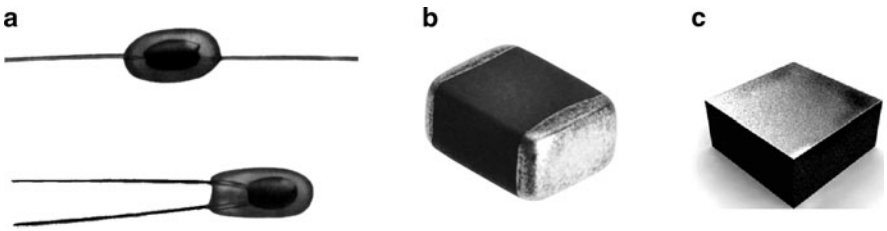

Fig. 16.11 Glass coated axial and radial bead thermistors (**a**); surface mounted thermistor (**b**); and top-bottom uncoated chip thermistor (**c**)

thermistors. The beads may be bare or coated with glass (Fig. 16.11), epoxy, or encapsulated into a metal jacket. All these beads have platinum alloy leadwires, which are sintered into the ceramic body. The platinum is selected because it combines a good electrical conductivity with not-so-good thermal conductivity. When fabricated, a small portion of mixed metal oxide with a suitable binder is placed onto a parallel leadwires, which are under slight tension. After the mixture has been allowed to dry, or has been partly sintered, the strand of beads is removed from the supporting fixture and placed for the final sintering into a tubular furnace. The metal oxide shrinks onto the leadwires during this firing process and forms an intimate electrical bond. Then, the beads are individually cut from the strand and are given an appropriate coating.

Another type of thermistor is a chip thermistor with surface contacts for the leadwires. Usually, the chips are fabricated by a tape casting process, with subsequent screen printing, spraying, painting, or vacuum metallization of the surface electrodes. The chips are either bladed or cut into desired geometry. If desirable, the chips can be ground to meet the required tolerances. The chips are given two electrodes either axially or top-bottom. Presently, many thermistor chips are fabricated in standard forms (0201, 0402, 0603, and 0805 in the U.S. or 0603, 1005, 1608, and 2012 – metric) for the surface-mounted assembly on circuit boards.

The third type of thermistors is fabricated by the deposition of semiconductive materials on a suitable substrate such as glass, alumina, silicon, etc. These thermistors are preferable for integrated sensors and for a special class of thermal infrared detectors. A typical method of fabrication is silk screening.

Among the metallized surface contact thermistors, flakes and uncoated chips are the least stable. A moderate stability may be obtained by epoxy coating. The bead type with leadwires sintered into the ceramic body permit operation at higher temperatures, up to 550°C. The metallized surface contact thermistors usually are rated up to 150°C. Whenever a fast response time is required, bead thermistors are preferable; however, they are more expensive than the chip type. Besides, the bead thermistors are more difficult to trim to a desired nominal value. Trimming is usually performed by mechanical grinding of a thermistor at a selected temperature (usually 25°C) to change its geometry and thus to bring its resistance to a specified value.

While using the NTC thermistors, one must not overlook possible sources of error. One of them is aging, which for the low-quality sensors may be as large as

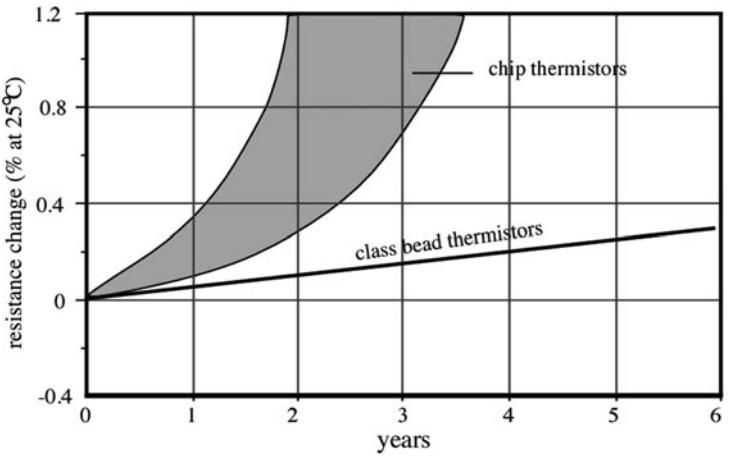

Fig. 16.12 Long-term stability of thermistors

+1%/year. Figure 16.12 shows typical percentage changes in resistance values for the epoxy-encapsulated chip thermistors when compared with the sintered glass encapsulated thermistors. A good environmental protection and preaging is a powerful method of sensor characteristic stabilizing. During preaging, the thermistor is maintained at +300°C for at least 700 h. For the better protection, it may be further encapsulated into a stainless steel jacket and potted with epoxy. After preaging and encapsulation in glass and then into a stainless steel tube, a thermistor may have a drift as low as few millidegree per year.

16.3.3.6 Germanium and Silicon NTC Thermistors

High-quality NTC thermistors may be successfully fabricated from monocrystalline or polycrystalline germanium or silicon [9, 10]. These thermistors have several advantages, some of them are very high sensitivity (typical β exceeds 6,000 K at 25°C), tight manufacturing tolerances, very small sizes, low cost, and ability to operate at cryogenic and high temperatures, from as low as 1 mK to as high as 500°C. Currently, these sensors are produced by AdSem, Inc. (www.adsem.com).

16.3.3.7 Printed Thermistors

A thermistor can be produced in a technique similar to conventional thick-film resistor [31]. The process involves screen-printing on a ceramic substrate. The ink for printing contains a powder with thermistor characteristics, glass powder, and organic binder. The thermistor powder is composed of oxides of Mn, Co, Ni, oxides of some noble metals such as Ru and other materials [11]. After printing, the

thermistors fired in a furnace (sintered), then contact electrodes are printed at the edges of the thermistor pattern and fired again. Then, the substrates are cut into individual thermistors. Currently, high-quality thermistor inks (pastes) are commercially available from DuPont. Examples are pastes NTC40 and NTC50.

16.3.3.8 Self-Heating Effect in NTC Thermistors

As it was mentioned earlier, when employing an NTC thermistor, a self-heating effect should not be overlooked. A thermistor is an active type of a sensor, that is, it does require an excitation signal for its operation. The signal is usually either a dc or ac passing through the thermistor. The electric current causes a Joule heating and a subsequent increase in temperature. In some applications, this may be a source of error that may result in an erroneous determination of temperature of a measured object. In other applications, the self-heating is successfully employed for sensing fluid flow, thermal radiation, and other stimuli.

Let us analyze the thermal events in a thermistor, when electric power is applied. Figure 16.13a shows a voltage source E connected to a thermistor R_T thorough a current limiting resistor R.

When electric power P is applied to the circuit (moment *on* in Fig. 16.13b), the rate at which energy is supplied to the thermistor must be equal to the rate at which energy H_L is lost plus the rate at which energy H_s is absorbed by the thermistor body. The absorbed energy is stored in the thermistor's thermal capacity C. The power balance equation is

$$\frac{dH}{dt} = \frac{dH_{\mathrm{L}}}{dt} + \frac{dH_{\mathrm{s}}}{dt}. \tag{16.30}$$

According to the law of conservation of energy, the rate at which thermal energy is supplied to the thermistor is equal to electric power delivered by voltage source E

$$\frac{dH}{dt} = P = \frac{V_T^2}{R_T} = V_T i, \tag{16.31}$$

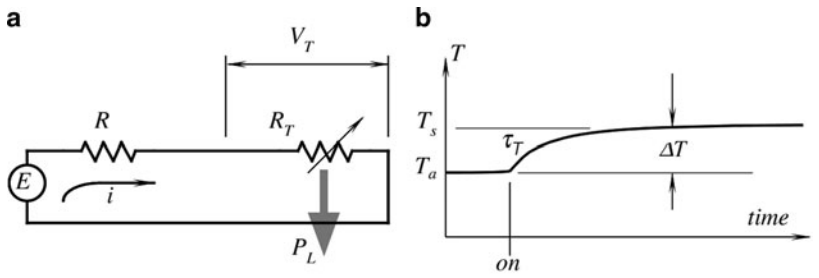

Fig. 16.13 Current passing through thermistor causes self-heating (**a**) and temperature of thermistor rises with thermal time constant τ_T. P_L is thermal power lost to surroundings (**b**)

where V_T is the voltage drop across the thermistor. The rate at which thermal energy is lost from the thermistor to its surroundings is proportional to temperature gradient ΔT between the thermistor and the surrounding temperature T_a:

$$P_L = \frac{dH_L}{dt} = \delta \Delta T = \Delta(T_s - T_a), \tag{16.32}$$

where δ is the so-called dissipation factor, which is equivalent to a thermal conductivity from the thermistor to its surroundings. It is defined as a ratio of dissipated power and a temperature gradient (at a given surrounding temperature). The factor depends upon the sensor design, length, and thickness of leadwires, thermistor material, supporting components, thermal radiation from the thermistor surface, and relative motion of medium in which the thermistor is located.

The rate of heat absorption is proportional to thermal capacity of the sensor assembly

$$\frac{dH_s}{dt} = C\frac{dT_s}{dt}. \tag{16.33}$$

This rate causes the thermistor's temperature T_s to rise above its surroundings. Substituting (16.32) and (16.33) into (16.31) we arrive at

$$\frac{dH}{dt} = P = Ei = \delta(T_s - T_a) + C\frac{dT_s}{dt}. \tag{16.34}$$

The above is a differential equation describing the thermal behavior of the thermistor. Let us now solve it for two conditions. The first condition is the constant electric power supplied to the sensor: $P = \text{const}$. Then, solution of (16.34) is

$$\Delta T = (T_s - T_a) = \frac{P}{\delta}\left[1 - e^{-\frac{\delta}{C}t}\right], \tag{16.35}$$

where e is the base of natural logarithms. The above solution indicates that upon applying electric power, the temperature of the sensor will exponentially rise above ambient. This specifies a transient condition, which is characterized by a thermal time constant $\tau_T = C(1/\delta)$. Here, the value of $1/\delta = r_T$ has a meaning of thermal resistance between the sensor and its surroundings. The exponential transient is shown in Fig. 16.13b.

Upon waiting sufficiently long to reach a steady-state level T_s, the rate of change in (16.34) becomes equal to zero ($dT_s/dt = 0$); then, the rate of heat loss is equal to supplied power

$$\delta(T_\mathrm{s} - T_\mathrm{a}) = \delta\Delta T = V_T i. \tag{16.36}$$

If by selecting low-supply voltage and high resistances, the current i is made very low, temperature rise, ΔT, can be made negligibly small and the self-heating is virtually eliminated. Then, from (16.34) becomes

$$\frac{dT_s}{dt} = -\frac{\delta}{C}(T_s - T_a). \tag{16.37}$$

The solution of this differential equation yields an exponential function (16.8), which means that the sensor responds to change in environmental temperature with time constant τ_T. Since the time constant depends on the sensor's coupling to the surroundings, it is usually specified for certain conditions, for instance, $\tau_T = 1$ s at 25°C, in still air or 0.1 s at 25°C in stirred water. It should be kept in mind that the above analysis represents a simplified model of the heat flows. In the reality, a thermistor response has a somewhat nonexponential shape.

All thermistor applications require the use of one of three basic characteristics:

1. The resistance versus temperature characteristic of the NTC thermistor is shown in Fig. 16.14. In most of the applications based on this characteristic, the self-heating effect is undesirable. Thus, the nominal resistance R_{T0} of the thermistor should be selected high and its coupling to the object should be maximized (increase in δ). The characteristic is primarily used for sensing and measuring temperature. Typical applications are contact electronic thermometers, thermostats, and thermal breakers.
2. The current versus time (or resistance versus time), like the one shown in Fig. 16.12b.
3. The voltage versus current characteristic is important for applications where the self-heating effect is employed, or otherwise can not be neglected. The power supply–loss balance is governed by (16.36). If variations in δ are small (which is often the case) and the resistance versus temperature characteristic is known, then (16.36) can be solved for the static voltage versus current characteristic.

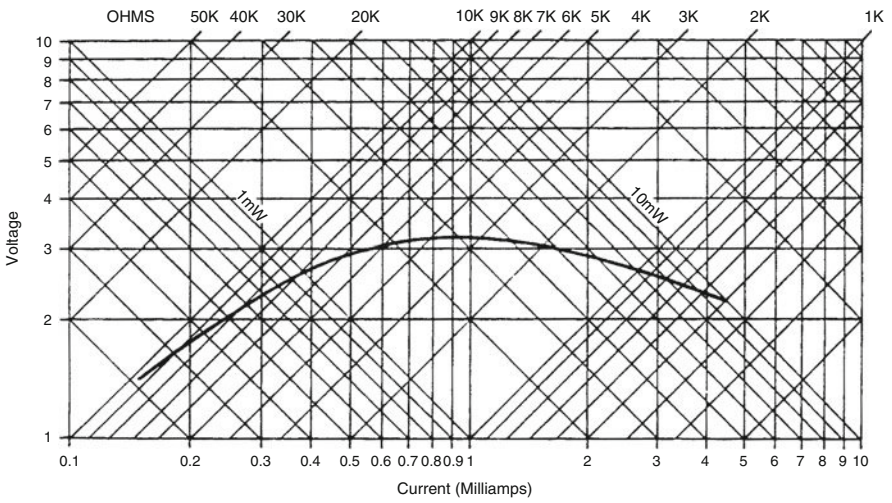

Fig. 16.14 Voltage–current characteristic of an NTC thermistor in still air at 25°C (curvature of the characteristic is due to a self-heating effect)

That characteristic is usually plotted on log–log coordinates, where lines of constant resistance have a slope of +1 and lines of constant power have slope of −1 (Fig. 16.14).

At very low currents (left side of the Fig. 16.14), the power dissipated by the thermistor is negligibly small, and the characteristic is tangential to a line of constant resistance of the thermistor at a specified temperature. Thus, the thermistor behaves as a simple resistor. That is, voltage drop V_T is proportional to current i.

As the current increases, the self-heating increases as well. This results in a decrease in the resistance of the thermistor. Since the resistance of the thermistor is no longer constant, the characteristics start to depart from the straight line. The slope of the characteristic (dV_T/di), which is the resistance, drops with increase in current. The current increase leads to further resistance drop, which, in turn, increases the current. Eventually, current will reach its maximum value i_p at a voltage maximum value V_p. It should be noted that, at this point, a resistance of the thermistor is zero. Further increase in current i_p will result in continuing decrease in the slope, which means that the resistance has a negative value (right side of Fig. 16.14). An even further increase in current will produce another reduction in resistance, where leadwire resistance becomes a contributing factor. A thermistor should never be operated under such conditions. A thermistor manufacturer usually specifies the maximum power rating for thermistors.

According to (16.36), self-heating thermistors can be used to measure variations in δ, ΔT, or V_T. The applications where δ varies include vacuum manometers (Pirani gauges), anemometers, flow meters, fluid level sensors, etc. Applications where ΔT is the stimulus include microwave power meters, AFIR detectors, etc. The applications where V_T varies are in some electronic circuits: Automatic gain control, voltage regulation, volume limiting, etc.

16.3.3.9 PTC Thermistors

All metals may be called PTC materials. Their temperature coefficients of resistivity (TCR) are quite low (Table A.7), thus making them not very useful for temperature sensing that require high sensitivity. The RTDs as described earlier also have small PTCs. In contrast, ceramic PTC materials in a certain temperature range are characterized by a very large temperature dependence. The PTC thermistors are fabricated of polycrystalline ceramic substances, where the base compounds, usually barium titanate or solid solutions of barium and strontium titanate (highly resistive materials) made semiconductive by the addition of dopants [12]. Above the Curie temperature of a composite material, the ferroelectric properties change, rapidly resulting in a rise in resistance, often several orders of magnitude. A typical transfer function curve for the PTC thermistor is shown in Fig. 16.15 in comparison with the NTC and platinum RTD responses. The shape of the curve does not lend itself to an easy mathematical approximation; therefore, manufacturers usually specify PTC thermistors by a set of numbers:

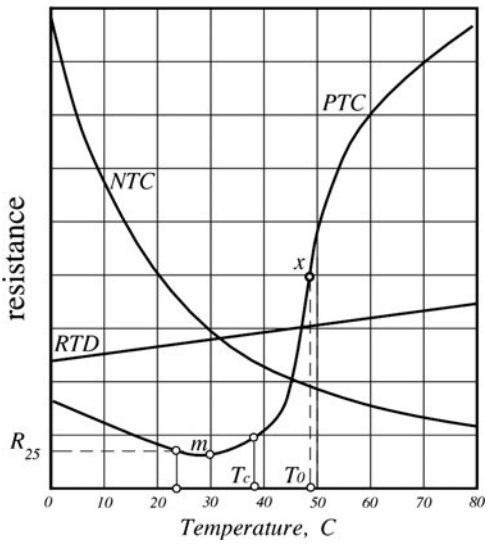

Fig. 16.15 Transfer functions of PTC and NTC thermistors as compared with RTD

1. Zero-power resistance, R_{25}, at 25°C, where self-heating is negligibly small.
2. Minimum resistance R_m is the value on the curve where thermistor changes its TCR from positive to negative value (point m).
3. Transition temperature T_τ is the temperature where resistance begins to change rapidly. It coincides approximately with the Curie point of the material. A typical range for the transition temperatures is from −30 to +160°C (Keystone Carbon Co.).
4. TCR is defined in a standard form.

$$\alpha = \frac{1}{R}\frac{\Delta R}{\Delta T}. \tag{16.38}$$

The coefficient changes very significantly with temperature and often is specified at point x, that is, at its highest value, which may be as large as 2/°C (meaning the change in resistance is 200% per °C).

5. Maximum voltage $E_{\max}$ is the highest value that the thermistor can withstand at any temperature.
6. Thermal characteristics are specified by a thermal capacity, a dissipation constant δ (specified under given conditions of coupling to the environment), and a thermal time constant (defines speed response under specified conditions).

It is important to understand that for the PTC thermistors, two factors play a key role: environmental temperature and a self-heating effect. Either one of these two factors shifts the thermistor's operating point.

The temperature sensitivity of the PTC thermistor is reflected in a volt–ampere characteristic of Fig. 16.16. A regular resistor with the near zero TCR, according to Ohm's law, has a linear characteristic. A NTC thermistor has a positive curvature of the volt–ampere dependence. An implication of the negative TCR is that if such a thermistor

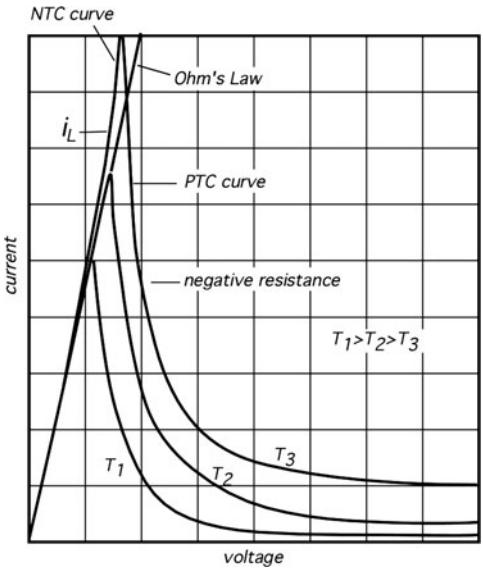

Fig. 16.16 Volt–ampere characteristic of a PTC thermistor

is connected to a hard voltage source,[7] a self-heating due to Joule heat dissipation will result in resistance reduction. In turn, that will lead to further increase in current and more heating. If the heat outflow from the NTC thermistor is restricted, a self-heating may eventually cause overheating and a catastrophic destruction of the device.

Thanks to positive TCRs, metals do not overheat when connected to hard voltage sources and behave as self-limiting devices. For instance, a filament in an incandescent lamp does not burn out because the increase in its temperature results in an increase in resistance, which limits current. This self-limiting (self-regulating) effect is substantially enhanced in the PTC thermistors. The shape of the volt–ampere characteristic indicates that in a relatively narrow temperature range, the PTC thermistor possesses a negative resistance, that is

$$R_x = -\frac{dV_x}{di}. \tag{16.39}$$

This results in the creation of an internal negative feedback that makes this device a self-regulating thermostat. In the region of a negative resistance, any increase in voltage across the thermistor results in heat production, which, in turn, increases the resistance and reduces heat production. As a result, the self-heating effect in a PTC thermistor produces enough heat to balance the heat loss on such a level that it

[7] A hard voltage source means any voltage source having a near zero output resistance and capable of delivering unlimited current without a change in voltage.

maintains the device's temperature on a constant level T_0 (Fig. 16.15). That temperature corresponds to point x where tangent to the curve has the highest value.

It should be noted that PTC thermistors are much more efficient when T_0 is relatively high (over 100°C) and their efficiency (the slope of the R–T curve near point x) drops significantly at lower temperatures. By their very nature, PTC thermistors are useful in the temperature range that is substantially higher than the operating ambient temperature.

Even though the PTC thermistors are not appropriate for accurate measurement of temperature, there are several applications where the self-regulating effect of a PTC thermistor may be quite useful. We briefly mention three of them:

1. Circuit protection: A PTC thermistor may operate as a nondestructible (resettable) fuse in electric circuits, sensing excessive currents. Figure 16.17a shows a PTC thermistor connected in series with a power supply voltage E feeding the load with current i. The resistance of the thermistor at room temperature is quite low (typically from 10 to 140 Ω). Current i develops voltage V_L across the load and voltage V_x across the thermistor. It is assumed that $V_L \gg V_x$. Power dissipated by the thermistor $P = V_x i$, is lost to the surroundings and the thermistor's temperature is raised above ambient by a relatively small value. Whenever either ambient temperature becomes too hot or load current increases dramatically (for instance, due to internal failure in the load), the heat dissipated by the thermistor elevates its temperature to a T_τ region where its resistance starts increasing. This limits further current increase. Under the shorted-load conditions, $V_x = E$ and current i drops to its minimal level. This will be maintained until normal resistance of the load is restored and, it is said, that the fuse resets itself. It is important to assure that $E < 0.9E_{\max}$; otherwise, a catastrophic destruction of the thermistor may occur.
2. A miniature self-heating thermostat (Fig. 16.17b) for the microelectronic, biomedical, chemical, and other suitable applications can be designed with a single PTC thermistor. Its transition temperature must be appropriately selected. A thermostat consists of a dish, which is thermally insulated from the environment and thermally coupled to the thermistor. Thermal grease is recommended to

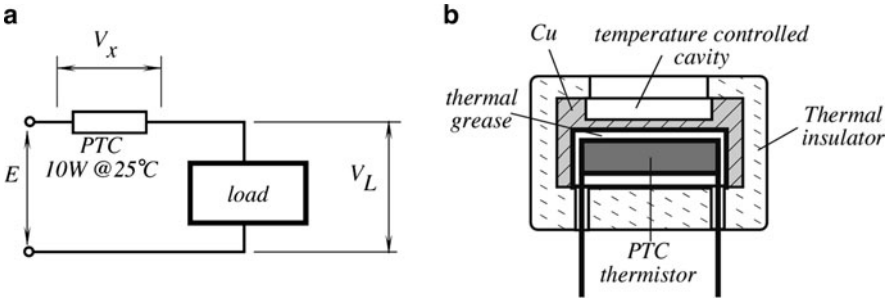

Fig. 16.17 Applications of PTC thermistors: Current limiting circuit (**a**) and mini-thermostat (**b**)

eliminate a dry contact. The terminals of the thermistor are connected to a voltage source whose value may be estimated from the following formula

$$E \geq 2\sqrt{\delta(T_\tau - T_a)R_{25}}, \tag{16.40}$$

where δ is the heat dissipation constant, which depends on thermal coupling to the environment and T_a is the ambient temperature. The thermostat's set point is determined by the physical properties of the ceramic material (Curie temperature) and due to internal thermal feedback, the device reliably operates within relatively large range of power supply voltages and ambient temperatures. Naturally, ambient temperature must be always less than T_τ.

3. Time delay circuits can be created with the PTC thermistors thanks to a relatively long transition time between the application of electric power in its heating and a low-resistance point.
4. Flowmeter and liquid-level detectors that operate on principle of heat dissipation (thermo-anemometers as described in Sect. 11.3) can be made very simple with the PTC thermistors.

16.4 Thermoelectric Contact Sensors

A thermoelectric contact sensor is called a *thermocouple* because at least two dissimilar conductors are joined to form a junction (a couple). However, at least two of these junctions are needed to make a practical sensor. A thermocouple is a passive sensor. It generates voltage in response to temperature and does not require any external excitation power. In other words, a thermocouple is a direct converter of thermal energy into electrical energy and because it is a voltage-generating sensor, sometimes thermocouple is called a "thermal battery."

The thermoelectric sensors belong to class of the *relative* sensors, because the voltage produced depends on a temperature difference between two thermocouple junctions, in large part regardless of the absolute temperature of each junction. To measure temperature with a thermocouple, one junction will serve as a reference and its absolute temperature must be measured by a separate absolute sensor such a thermistor, RTD, etc. or coupled to a material that is in a state of a known reference temperature (Table 16.1). Section 3.9 provides a physical background for better understanding of the thermoelectric effect, and Table A.10 lists some popular thermocouples[8] that are designated by letters originally assigned by the Instrument Society of America (ISA) and adopted by an American Standard in ANSI MC 96.1. Detailed description of various thermocouples and their applications can be found in many excellent texts, for instance in [2, 13, 14]. In the following paragraphs, we summarize the most important recommendations for the use of these sensors.

[8]A lot of very valuable information can be found in www.temperatures.com

Type T: Cu (+) versus constantan (−), are resistant to corrosion in moist atmosphere and are suitable for subzero temperature measurements. Their use in air in oxidizing environment is restricted to 370°C (700°F) due to the oxidation of the copper thermoelement. They may be used to higher temperatures in some other atmospheres.

Type J: Fe (+) versus constantan (−) are suitable in vacuum and in oxidizing, reducing, or inert atmospheres, over the temperature range of 0 to 760°C (32–1,400°F). The rate of oxidation in the iron thermoelement is rapid above 540°C (1,000°F), and the use of heavy-gage wires is recommended when long life is required at the higher temperatures. This thermocouple is not recommended for use below the ice point because rusting and embrittlement of the iron thermoelement make its use less desirable than Type T.

Type E: 10%Ni/Cr (+) versus constantan (−), are recommended for use over the temperature range of −200 to 900°C (−330 to 1,600°F) in oxidizing or inert atmospheres. In reducing atmospheres, alternately oxidizing or reducing atmospheres, marginally oxidizing atmospheres, and in vacuum, they are subject to the same limitations as type K. These thermocouples are suitable to subzero measurements since they are not subject to corrosion in atmospheres with a high-moisture content. They develop the highest e.m.f. per degree of all the commonly used types and are often used primarily because of this feature (see Fig. 3.36).

Type K: 10%Ni/Cr (+) versus 5%Ni/Al/Si (−), are recommended for use in an oxidizing or completely inert atmosphere over a temperature range of −200 to 1,260°C (−330 to 2,300°F). Due to their resistance to oxidation, they are often used at temperatures above 540°C. However, type K should not be used in reducing atmospheres, in sulfurous atmospheres, and in a vacuum.

Types R and S: Pt/Rh (+) versus Pt (−), are recommended for continuous use in oxidizing or inert atmospheres over a temperature range of 0–1,480°C (32–2,700°F).

Type B: 30%Pt/Rh (+) versus 6%Pt/Rh (−) are recommended for continuous use in oxidizing or inert atmospheres over the range of 870–1,700°C (1,000–3,100°F). They are also suitable for short term use in a vacuum. They should not be used in reducing atmospheres, nor those containing metallic or nonmetallic vapors. They should never be directly inserted into a metallic primary protecting tube or well.

16.4.1 Thermoelectric Laws

For practical purposes, an application engineer must be concerned with three basic laws that establish the fundamental rules for proper connection of the thermocouples. It should be stressed, however, that an electronic interface circuit must always be connected to two *identical* conductors; otherwise, two parasitic thermocouples will be formed at the circuit. These identical conductors may be formed from one of the thermocouple loop arms. That arm is broken to connect the metering device to the circuit. The broken arm is indicated as material *A* in Fig. 16.18.

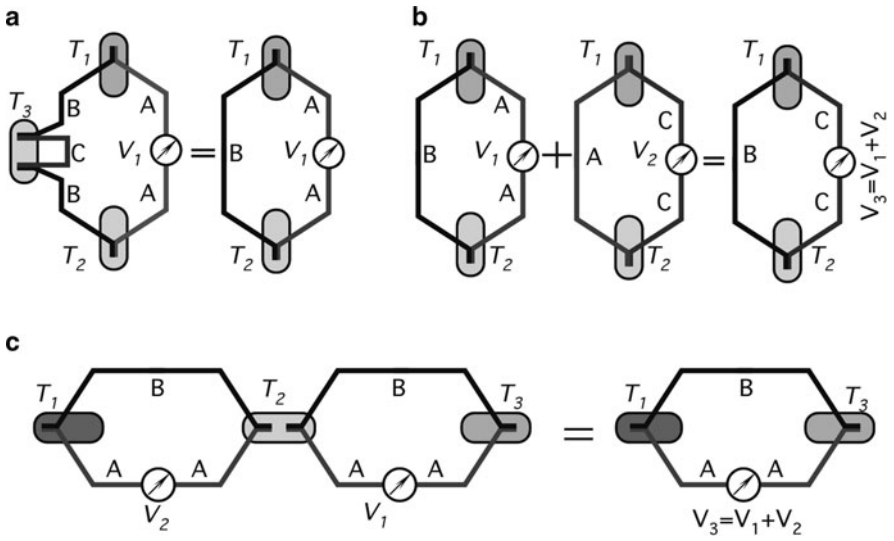

Fig. 16.18 Illustrations for the laws of thermocouples

Law No. 1: A thermoelectric current cannot be established in a homogeneous circuit by heat alone.

This law provides that a nonhomogeneous material is required for the generation of the Seebeck potential. If a conductor is homogeneous, regardless of the temperature distribution along its length, the resulting voltage is zero. The junction of two dissimilar conductors provide a condition for voltage generation.

Law No. 2: The algebraic sum of the thermoelectric forces in a circuit composed of any number and combination of dissimilar materials is zero if all junctions are at a uniform temperature.

The law provides that an additional material C, which can be inserted into any arm of the thermoelectric loop without affecting the resulting voltage V_1 as long as both additional joints are at the same temperature (T_3 in Fig. 16.18a). There is no limitation on the number of inserted conductors, as long as both contacts for each insertion are at the same temperature. This implies that an interface circuit must be attached in such a manner as to assure a uniform temperature for both contacts. Another consequence of the law is that thermoelectric joints may be formed by any technique, even if an additional intermediate material is involved (for instance, solder). The joints may be formed by welding, soldering, twisting, fusion, and so on without affecting the accuracy of the Seebeck voltage. The law also provides a rule of additive materials (Fig. 16.18b): If thermoelectric voltages (V_1 and V_2) of two conductors (B and C) with respect to a reference conductor (A) are known, the voltage of a combination of these two conductors is the algebraic sum of their voltages against the reference conductor.

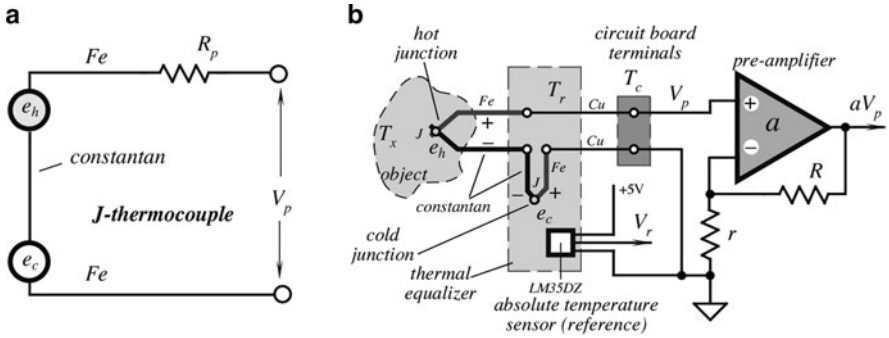

Fig. 16.19 Use of a thermocouple equivalent circuit of a thermocouple (**a**) and front end of a thermometer with a semiconductor reference sensor (**b**)

Law No. 3: If two junctions at temperatures T_1 and T_2 produce Seebeck voltage V_2, and temperatures T_2 and T_3 produce voltage V_1, then temperatures T_1 and T_3 will produce $V_3 = V_1 + V_2$ (Fig. 16.18c). This is sometimes called the *law of intermediate temperatures.*

The law allows us to calibrate a thermocouple at one temperature interval and then to use it at another interval. It also provides that extension wires of the same combination may be inserted into the loop without affecting the accuracy.

The abovementioned laws provide for numerous practical circuits where thermocouples can be used in a great variety of combinations. They can be arranged to measure the average temperature of an object, to measure the differential temperature between two objects, and to use other than thermocouple sensors for the reference junctions, etc.

It should be noted that thermoelectric voltage is quite small, and the sensors, especially with long connecting wires, are susceptible to various transmitted interferences. To increase the output signal, several thermocouples may be connected in series, while all reference junctions and all measuring junctions are maintained at the respective temperatures. Such an arrangement is called a *thermopile* (like piling up several thermocouples). Traditionally, the reference junctions are called cold and the measuring junctions are called hot.

Figure 16.19a shows an equivalent circuit for a thermocouple and a thermopile. Each junction consists of a voltage source and a serial resistor. The voltage sources represent the hot (e_h) and cold (e_c), while the resistances are combined in a resistor R_p. Seebeck potentials and the net voltage V_p have magnitude, which is the function of a temperature differential. The terminals of the circuit are assumed being fabricated of the same material, iron in this example.

16.4.2 Thermocouple Circuits

In the past, thermocouples were often used with a cold junction immersed into a reference melting ice bath to maintain its temperature at 0°C (thus, the "cold"

junction name). This presents serious limitations for many practical uses. The second and third thermoelectric laws provided earlier allow for a simplified solution. A "cold" junction can be maintained at any temperature, including ambient, as long as that temperature is precisely known. Therefore, a "cold" junction is thermally coupled to an additional temperature sensor, which does not require reference compensation. Usually, such a sensor is either thermoresistive or a semiconductor.

Figure 16.16b shows the correct connection of a thermocouple to an electronic circuit. Both the "cold" junction and the reference sensor must be positioned in an intimate thermal coupling. Often, they are imbedded into a chunk of copper or aluminum. To avoid dry contact, thermally conductive grease or epoxy should be applied for better thermal tracking. A reference temperature detector in this example is a semiconductor sensor LM35DZ manufactured by National Semiconductor, Inc. The circuit has two outputs – one for the signal representing the Seebeck voltage V_p and the other for the reference signal V_r. The schematic illustrates that connections to the circuit board input terminals and then to the amplifier's non-inverting input and to the ground bus is made by the same type of wires (Cu). Both board terminals should be at the same temperature T_c; however, they not necessarily have to be at the "cold" junction temperature. This is especially important for the remote measurements, where circuit board temperature may be different from the reference "cold" junction temperature T_r.

For computing the absolute temperature from a thermocouple sensor, two signals are essentially required. The first is a thermocouple voltage V_p and the other is the reference sensor output voltage V_r. These two signals come from different types of sensors and therefore are characterized by different transfer functions. A thermopile in most cases may be considered linear with normalized sensitivity α_p [V/K], while the reference sensor sensitivity is expressed according to its nature. Foe example, a thermistor's sensitivity α_r at the operating temperature T is governed by (16.21) and has dimension [Ω/K]. There are several practical ways of processing the output signals. One and the most precise is to measure these signals separately, then compute the reference temperature T_r according to the reference sensor's equation and compute the gradient temperature Δ from a thermocouple voltage V_p, as

$$\Delta = T_x - T_r = \frac{V_p}{\alpha_p}. \tag{16.41}$$

Finally, add the two temperatures Δ and T_p to arrive at the measured absolute temperature T_x. A value of sensitivity (α_p) can be found from Table A.10.

For a relatively narrow reference temperature range, instead of adding up the temperatures, voltages from the reference sensor and the thermocouple can be combined instead. Since α_r and α_p are very much different, a scaling circuit must be employed. Figure 16.20 illustrates a concept of adding up voltages from a thermocouple and a thermistor (reference sensor) to obtain a combined outputs signal V_c. When adding up the voltages, the thermocouple amplifier gain should be selected to match the temperature sensitivities of voltages V_p and V_r, or in other words, to satisfy condition

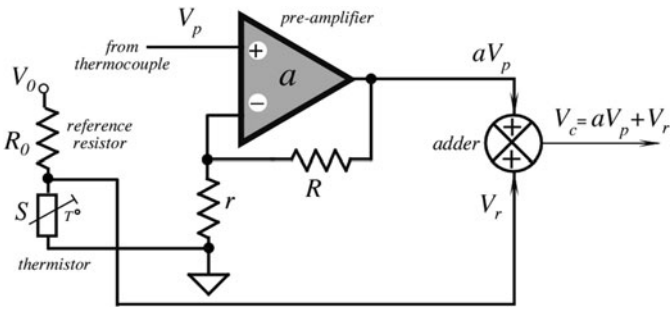

Fig. 16.20 Combining thermopile and thermistor signals

$$a\alpha_{\mathrm{p}} = \alpha_{\mathrm{r}} \tag{16.42}$$

It is preferable to select $R_0 = S_0$ (S_0 is the thermistor resistance at the calibrating temperature T_0 (in kelvin), for example, at $T_0 = 298.15$ K (25°C) or in the middle of the operating temperature range). With (16.21) in mind, and after differentiating voltage V_r, we arrive at the amplifier's gain

$$a = \frac{V_0}{\alpha_p}\frac{\beta}{T_0^2}\frac{R_0 S_0}{(R_0 + S_0)^2} \approx \frac{V_0}{4\alpha_p}\frac{\beta}{T_0^2}, \tag{16.43}$$

where V_0 is a constant voltage and β is the thermistor's characteristic temperature. The measured temperature can be computed from one of the corresponding equations found in Table 16.2, depending on the thermistor's model used. When a particular thermistor model is selected, temperature is computed from a virtual thermistor's resistance S_{c} that first is derived from the combined voltage V_{c} as

$$S_c = R_0 \frac{V_c}{V_0 - V_c}. \tag{16.44}$$

16.4.3 Thermocouple Assemblies

A complete thermocouple-sensing assembly generally consists of one or more of the following: a sensing element assembly (the junction), a protective tube (ceramic or metal jackets), a thermowell (for some critical applications, these are drilled solid bar stocks, which are made to precise tolerances and are highly polished to inhibit corrosion), terminations (contacts that may be in the form of a screw type, open type, plug and jack-disconnect, military standard type connectors, etc.). Some typical thermocouple assemblies are shown in Fig. 16.21. The wires may be left bare or given electrical isolators. For the high-temperature applications, the isolators may be of a

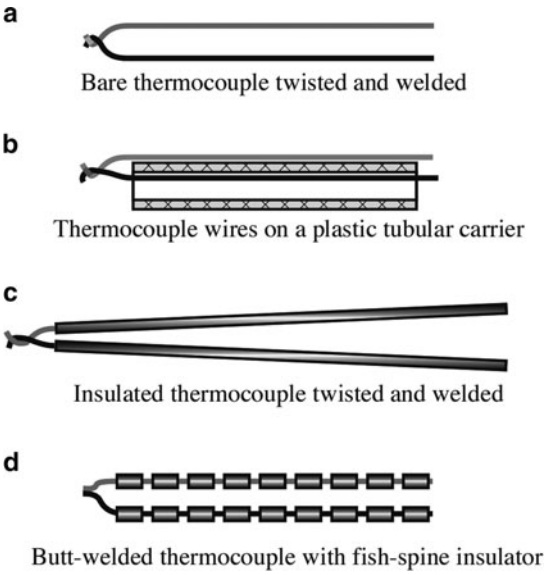

Fig. 16.21 Some thermocouple assemblies

fish-spine or ball ceramic type, which provides sufficient flexibility. If thermocouple wires are not electrically isolated, a measurement error may occur. Insulation is affected adversely by moisture, abrasion, flexing, temperature extremes, chemical attack, and nuclear radiation. A good knowledge of particular limitations of insulating materials is essential for accurate and reliable measurement. Some insulations have a natural moisture resistance. Teflon, polyvinyl chloride (PVC), and some forms of polyimides are examples of this group. With the fiber-type insulations, moisture protection results from impregnating with substances such as wax, resins, or silicone compounds. It should be noted that only one-time exposure to overextreme temperatures cause evaporation of the impregnating materials and loss of protection.

The moisture penetration is not confined to the sensing end of the assembly. For example, if a thermocouple passes through hot or cold zones, condensation may produce errors in the measurement, unless adequate moisture protection is provided.

The basic types of flexible insulations for elevated temperature usage are fiber glass, fibrous silica, and asbestos (which should be used with proper precaution due to health hazard). In addition, thermocouples must be protected from atmospheres that are not compatible with the alloys. Protecting tubes serve the double purpose of guarding the thermocouple against mechanical damage and interposing a shield between the wires and the environment. The protecting tubes can be made of carbon steels (up to 540°C in oxidizing atmospheres), stainless steel (up to 870°C), ferric stainless steel (AISI 400 series), high-nickel alloys, Nichrome,[9] Inconel,[10] etc. (up to 1,150°C in oxidizing atmospheres).

[9]Trademark of the Driver-Harris Company.

[10]Trademark of the International Nickel Company.

Practically all base–metal thermocouple wires are annealed or given a "stabilizing heat treatment" by the manufacturer. Such treatment generally is considered sufficient, and seldom it is found advisable to further anneal the wire before testing or using. Although a new platinum and platinum–rhodium thermocouple wire as sold by some manufacturers already is annealed, it has become a regular practice in many laboratories to anneal all types of R, S, and B thermocouples, whether new or previously used, before attempting an accurate calibration. This is accomplished usually by heating the thermocouple electrically in air. The entire thermocouple is supported between two binding posts, which should be close together, so that the tension in the wires and stretching while hot are kept at a minimum. The temperature of the wire is conveniently determined with an optical pyrometer. Most of the mechanical strains are relieved during the first few minutes of heating at 1,400–1,500°C.

Thin film thermocouples are formed by bonding junctions of foil metals. They are available in a free filament style with a removable carrier and in a matrix style with a sensor embedded in a thin, laminated material. The foil having a thickness in the order of 5 μm (0.0002″) gives an extremely low mass and thermal capacity. Thin flat junctions may provide intimate thermal coupling with the measured surface. Foil thermocouples are very fast (a typical thermal time constant is 10 ms) and can be used with any standard interface electronic apparatuses. While measuring temperature with sensors having small mass, thermal conduction through the connecting wires always must be accounted for (see Fig. 16.1). Thanks to a very large length to thickness ratio of the film thermocouples (on the order of 1,000) heat loss via wires usually is negligibly small.

To attach a film thermocouple to an object, several methods are generally used. Some of them are various cements and flame or plasma sprayed ceramic coatings. For ease of handling, the sensors often are supplied on a temporary carrier of polyimide film (e.g., Kapton, see Chap. 18) which is tough, flexible, and dimensionally stable. It is exceptionally heat resistant and inert. During the installation, the carrier can be easily pilled off or released by application of heat. The free foil sensors can be easily brushed into a thin layer to produce an ungrounded junction. While selecting cements, care must be taken to avoid corrosive compounds. For instance, cements containing phosphoric acid are not recommended for use with thermocouples having copper in one arm.

16.5 Semiconductor *pn*-Junction Sensors

A semiconductor pn-junction in a diode and a bipolar transistor exhibits quite a strong thermal dependence [15]. If the forward biased junction is connected to a constant current generator, the resulting voltage becomes a measure of the junction temperature (Fig. 16.23a). A very attractive feature of such a sensor is its high degree of linearity (Fig. 16.22). This allows a simple method of calibration using just two points to define a slope (sensitivity) and an intercept.

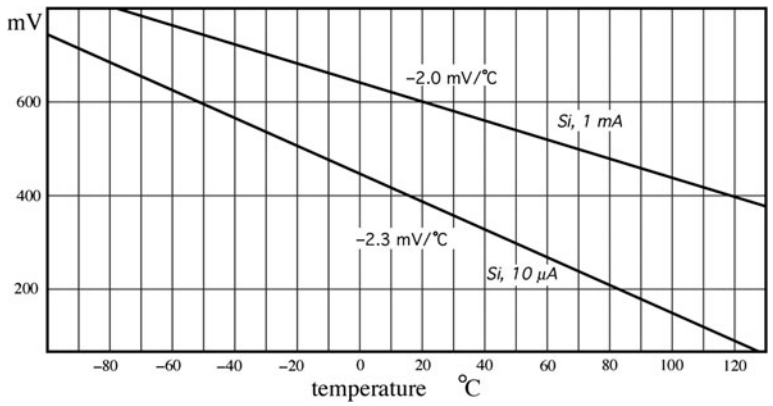

Fig. 16.22 Voltage-to-temperature dependence of a forward biased semiconductor junction under constant current conditions

The current-to-voltage equation of a pn-junction diode can be expressed as

$$I = I_0 e^{\frac{qV}{2kT}}, \tag{16.45}$$

where I_0 is the saturation current, which itself is a strong function of temperature. It can be shown that the temperature-dependent voltage across the junction can be expressed as

$$V = \frac{E_g}{q} - \frac{2kT}{q}(\ln K - \ln I), \tag{16.46}$$

where E_g is the energy band gap for silicon at 0 K (absolute zero), q is the charge of an electron, and K is a temperature independent constant. It follows from the above equation that when the junction is operated under constant current conditions, the voltage linearly relates to the temperature, and the slope is given by

$$b = \frac{dV}{dT} - \frac{2k}{q}(\ln K - \ln I). \tag{16.47}$$

Typically, for a silicon junction operating at 10 μA, the slope (sensitivity) is approximately −2.3 mV/°C and it drops to about −2.0 mV/°C for a 1 mA current. Any diode or junction transistor can be used as a temperature sensor. A practical circuit for the transistor used as a temperature sensor is shown in Fig. 16.23b. A voltage source E and a stable resistor R are used instead of a current source. Current through the transistor is determined as

$$I = \frac{E - V}{R}. \tag{16.48}$$

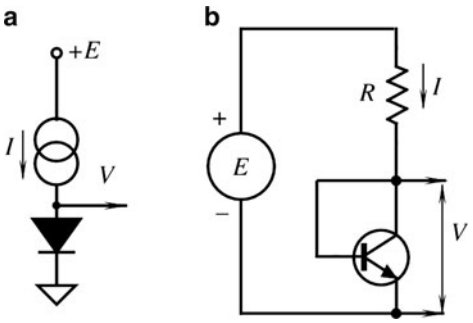

Fig. 16.23 Forward biased *pn*-junction temperature sensors Diode (**a**) and diode-connected transistor (**b**)

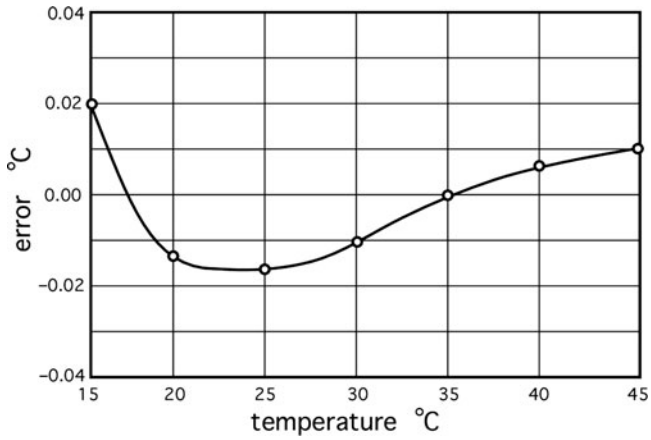

Fig. 16.24 An error curve for a silicon transistor (PN100) as a temperature sensor

It is recommended to use current on the order of $I = 100$ μA, therefore for $E = 5$ V and $V \approx 0.6$ V, the resistor $R = (E - V)/I = 44$ kΩ. When the temperature increases, voltage V drops, which results in a minute increase in current I. According to (16.47), this causes some reduction in sensitivity, which, in turn, is manifested as nonlinearity. However, the nonlinearity may be either small enough for a particular application, or it can be taken care of during the signal processing. This makes a transistor (diode) temperature sensor a very attractive device for many applications, due to its simplicity and very low cost. Figure 16.24 shows an error curve for the temperature sensors made with the PN100 transistor operating at 100 μA. It is seen that the error is quite small, and for many practical purposes, no linearity correction is required.

A diode sensor can be formed in a silicon substrate in many monolithic sensors that may require temperature compensation. For instance, it can be diffused into a micromachined membrane of a silicon pressure sensors to compensate for temperature dependence of piezoresistive elements.

An inexpensive, yet precise semiconductor temperature sensor may be fabricated by using fundamental properties of transistors to produce voltage that is

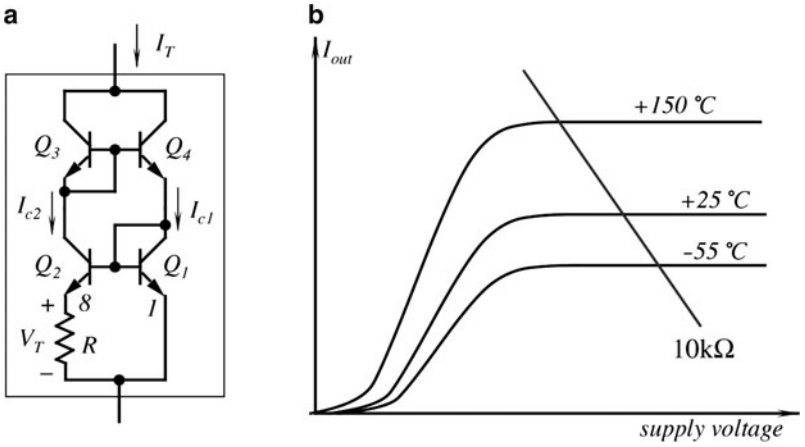

Fig. 16.25 Simplified circuit for a semiconductor temperature sensor (**a**) and current-to-voltage curves (**b**)

proportional to absolute temperature (in kelvin). That voltage can be used directly or it can be converted into current [16]. The relationship between base–emitter voltage (V_{be}) and collector current of a bipolar transistor is the key property to produce a linear semiconductor temperature sensor. Figure 16.25a shows a simplified circuit where Q_3 and Q_4 form the so-called current mirror. It forces two equal currents $I_{C1} = I$ and $I_{C2} = I$ into transistors Q_1 and Q_2. The collector currents are determined by resistor R. In a monolithic circuit, transistor Q_2 is actually made of several identical transistors connected in parallel, for example, 8. Therefore, the current density in Q_1 is eight times higher than that of each of transistors Q_2. The difference between base emitter voltages of Q_1 and Q_2 is

$$\Delta V_{be} = V_{be1} - V_{be2} = \frac{kT}{q} \ln\left(\frac{rI}{I_{ceo}}\right) - \frac{kT}{q} \ln\left(\frac{I}{I_{ceo}}\right) = \frac{kT}{q} \ln r, \tag{16.49}$$

where r is a current ratio (equal to 8 in our example), k is the Boltzmann constant, q is the charge of an electron, and T is the temperature in K. Currents I_{ceo} are the same for both transistors. As a result, a current across resistor R produces voltage $V_T = 179$ μV/T, which is independent on the collector currents. Therefore, the total current through the sensor is

$$I_T = 2\frac{V_T}{R} = \left(2\frac{k}{qR} \ln r\right) T \tag{16.50}$$

which for currents ratio $r = 8$, and resistor $R = 358$ Ω produces a linear transfer function $I_T/T = 1$ μA/°K.

Figure 16.25b shows current-to-voltage curves for different temperatures. Note that the value in the parenthesis of (16.50) is constant for a particular sensor design

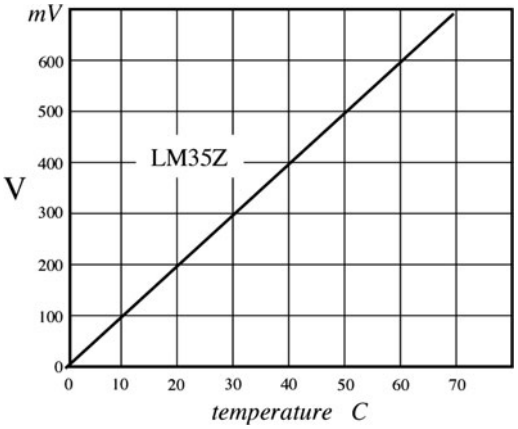

Fig. 16.26 A typical transfer function of a LM35DZ semiconductor temperature sensor (from National Semiconductors, Inc.)

and may be precisely trimmed during the manufacturing process for a desired slope I_T/T. The current I_T may be easily converted into voltage. If, for example, a 10 kΩ resistor is connected in series with the sensor, the voltage across that resistor will be a linear function of absolute temperature.

The simplified circuit of Fig. 16.25a will work according to the abovementioned equations only with the perfect transistors ($\beta = \infty$). Practical monolithic sensors contain many additional components to overcome limitations of the real transistors. Several companies produce temperature sensors based on this principle. Examples are LM35 from National Semiconductors (voltage output circuit) and AD590 from Analog Devices (current output circuit).

Figure 16.26 shows a transfer function of a LM35Z temperature sensor, which has a linear output internally trimmed for the Celsius scale with a sensitivity of 10 mV per °C. The function is quite linear where the nonlinearity error is confined within ± 0.1°. The function can be modeled by

$$V_{out} = V_0 + at, \tag{16.51}$$

where t is the temperature in degrees C. Ideally, V_0 should be equal to zero; however, part-to-part variations of its value may be as large as ±10 mV, which correspond to an error of 1°C. Slope a may vary between 9.9 and 10.1 mV/°C; hence, for a high-accuracy application, this sensor still may require a calibration to determine V_0 and a.

16.6 Optical Temperature Sensors

Temperature can be measured by contact and noncontact methods. The noncontact instruments are generally associated with the infrared optical sensors that we covered in Sects. 3.12.3 and 14.7. A need for the noncontact temperature sensors

exists when the measurement must be done fast. Also, they are needed for determining temperatures at hostile environments where very strong electrical, magnetic or electromagnetic fields, or very high voltages make measurements either too susceptible to interferences, or too dangerous for the operator. Also, there are situations when it is just difficult to reach an object during a routine measurement. Besides the infrared methods of temperature measurements, there are sensors that are contact by nature but still use photons as carriers of thermal information. These are discussed below.

16.6.1 Fluoroptic Sensors

The fluoroptic sensors rely on the ability of a special phosphor compound to give away a fluorescent signal in response to light excitation. The compound can be directly painted over the measured surface and illuminated by an ultraviolet (UV) pulse, while observing the afterglow. The shape of the response afterglow pulse is the function of temperature. The decay of the response afterglow is highly reproducible over a wide temperature range [17, 18]. As a sensing material, magnesium fluoromagnetite activated with tetravalent manganese is used. This is phosphor, long known in the lighting industry as a color corrector for mercury vapor street lamps, prepared as a powder by a solid-state reaction at approximately 1,200°C. It is thermally stable, relatively inert and benign from a biological standpoint, and insensitive to damage by most chemicals or by prolonged exposure to ultraviolet UV radiation. It can be excited to fluoresce by either UV or blue radiation. Its fluorescent emission is in the deep red region, and the fluorescent decay is essentially exponential.

To minimize crosstalk between the excitation and emission signals, they are passed through the bandpass filters, which reliably separate the related spectra (Fig. 16.27a). The pulsed excitation source, a xenon flash lamp, can be shared among a number of optical channels in a multisensor system. The temperature measurement is made by measuring the rate of decay of the fluorescence, as shown in Fig. 16.27b. In other words, a temperature is represented by a time constant τ, which drops fivefold over the temperature range from -200 to $+400$°C. Since measurement of time is usually the simplest and most precise operation that can be performed by an electronic circuit, the temperature can be measured with a good resolution and accuracy: About ± 2°C over the above range without calibration.

Since the time constant is independent of excitation intensity, a variety of designs is possible. For instance, the phosphor compound can be directly coated onto the surface of interest and the optic system can take measurement without a physical contact (Fig. 16.28a). This makes possible the continuous temperature monitoring without disturbing the measured site. In another design, a phosphor is coated on the tip of a pliable probe that can form a good contact area when touches the object (Fig. 16.28b, c).

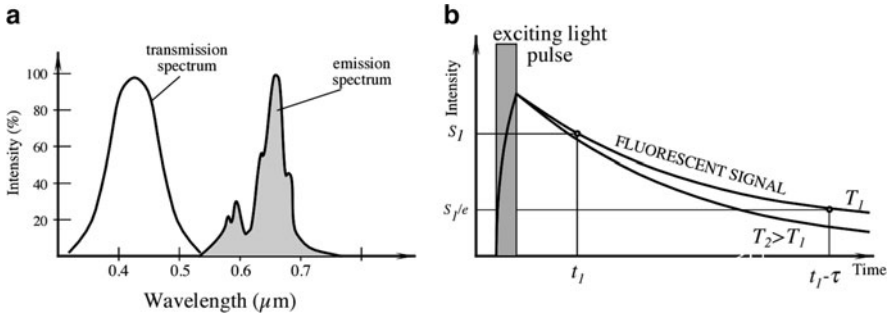

Fig. 16.27 Fluoroptic method of temperature measurement: Spectral responses of the excitation and emission signals (**a**) exponential decay of the emission signal for two temperatures (T_1 and T_2) (**b**); where e is the base of natural logarithms, and τ is a decay time constant (adapted from [17])

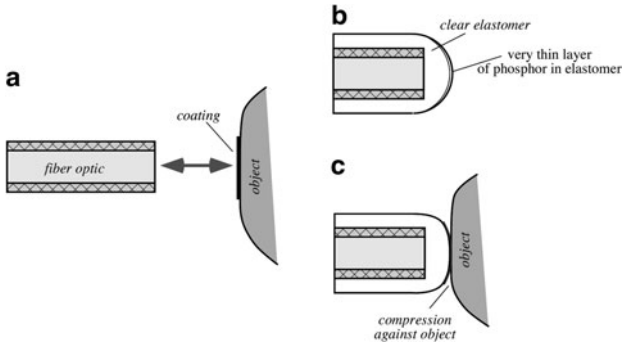

Fig. 16.28 Placement of a phosphor compound in fluoroptic method: On the surface of an object (**a**) and on the tip of the probe (**b** and **c**) (adapted from [17])

16.6.2 Interferometric Sensors

Another method of the optical temperature measurement is based on modulation of light intensity by interfering two light beams. One beam is a reference, while the other travel through a temperature-sensitive medium is somewhat delayed depending on temperature. This results in a phase shift and a subsequent extinction of the interference signal. For temperature measurement, a thin layer of silicon [19, 20] can be used because its refractive index changes with temperature, thus modulating a light travel distance.

Figure 16.29 shows a schematic of a thin film optical sensor. The sensor was fabricated by sputtering of three layers onto the ends of the step-index multimode fibers with 100 μm core diameters and 140 μm cladding diameters [21]. The first layer is silicon and then silicon dioxide. The FeCrAl layer on the end of the probe prevents oxidation of the underlying silicon. The fibers can be used up to 350°C;

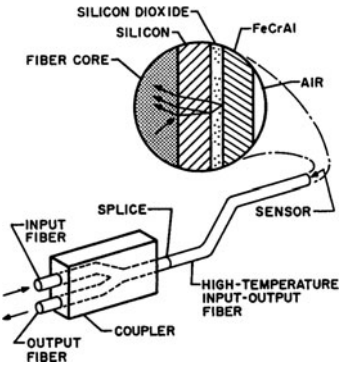

Fig. 16.29 A schematic of a thin film optical temperature sensor

however, much more expensive fibers with gold-buffered coatings can be used up to 650°C. The sensor is used with the LED light source operating in the range of 860 nm and a microoptic spectrometer.

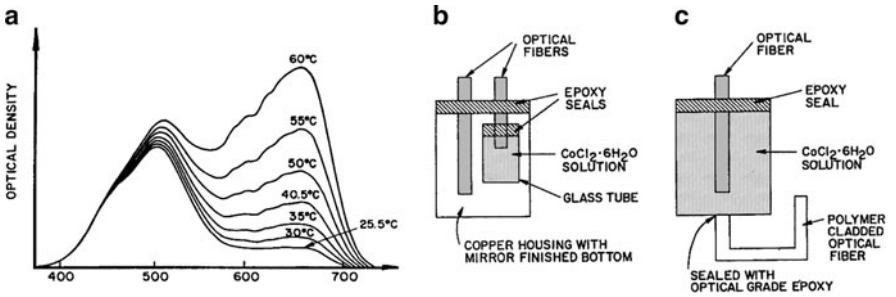

Fig. 16.30 A thermochromic solution sensor. Absorption spectra of the cobalt chloride solution (**a**); reflective fiber coupling (**b**); and transmissive coupling (**c**) (from [22])

16.6.3 Thermochromic Solution Sensor

For biomedical applications, where electromagnetic interferences may present a problem, a temperature sensor can be fabricated with the use of a thermochromic solution [22] such as cobalt chloride ($CoCl_2 \cdot 6H_2O$).

The operation of this sensor is based on the effect of a temperature dependence of a spectral absorption in the visible range of 400–800 nm by the thermochromic solution (Fig. 16.30a). This implies that the sensor should consist of a light source, a detector, and a cobalt chloride solution, which is thermally coupled with the object. Two possible designs are shown in Fig. 16.30b and c, where transmitting and receiving optical fibers are coupled through a cobalt chloride solution.

16.7 Acoustic Temperature Sensor

Under extreme conditions, temperature measurement may become a difficult task. These conditions include a cryogenic temperature range, high-radiation levels inside nuclear reactors, etc. Another unusual condition is the temperature measurement inside a sealed enclosure with a known medium, in which no contact sensors can be inserted and the enclosure in not transmissive for the infrared radiation. Under such unusual conditions, acoustic temperature sensors may come in quite handy. An operating principle of such a sensor is based on a relationship between temperature of the medium and speed of sound. For instance, in dry air at a normal atmospheric pressure, the relationship is approximated by

$$\upsilon \approx 331.5\sqrt{\frac{T}{273.15}}\ \mathrm{m/s}, \tag{16.52}$$

where v is the speed of sound, and T is the absolute temperature.

An acoustic temperature sensor (Fig. 16.31) is composed of three components: an ultrasonic transmitter, an ultrasonic receiver, and a gas-filled hermetically sealed tube. The transmitter and receiver are ceramic piezoelectric plates that are acoustically decoupled from the tube to assure sound propagation primarily through the enclosed gas, which in most practical cases is dry air. Alternatively, the transmitting and receiving crystals may be incorporated into a sealed enclosure with a known content whose temperature has to be measured. That is, an intermediate tube is not necessarily required in cases where the internal medium, its volume, and mass are held constant. When a tube is used, care should be taken to prevent its mechanical deformation and loss of hermeticity under the extreme temperature conditions. A suitable material for the tube is invar.

The clock of a low frequency (near 100 Hz) triggers the transmitter and disables the receiver, just like in a radar. The piezoelectric crystal flexes and that causes a transmission of an ultrasonic wave along the tube. The receiving crystal is enabled

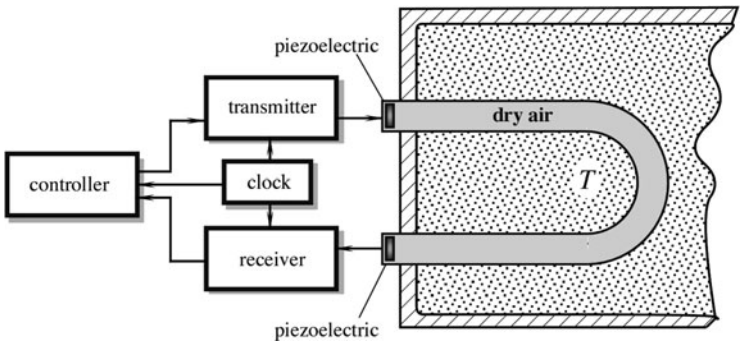

Fig. 16.31 Acoustic thermometer with an ultrasonic detection system

before the wave arrives to its surface and converts it into an electrical transient, which is amplified and sent to the control circuit. The control circuit calculates the speed of sound by determining propagation time along the tube. Then, the corresponding temperature is determined from the calibration numbers stored in a look-up table. In another design, the thermometer may contain only one ultrasonic crystal, which alternatively acts either as a transmitter or as a receiver. In that case, the tube has a sealed empty end. The ultrasonic waves are reflected from the end surface and propagate back to the crystal, which before the moment of the wave arrival is turned into a reception mode. An electronic circuit [23] converts the received pulses into a signal that corresponds to the tube temperature.

A miniature temperature sensor can be fabricated with the surface acoustic waves (SAW) and plate waves (PW) techniques (see Section 12.7). The idea behind such a sensor is in the temperature modulation of some mechanical parameters of a time-keeping element in the electronic oscillator [24, 25]. This leads to a change in the oscillating frequency. In effect, such an integral acoustic sensor becomes a direct converter of temperature into frequency. A typical sensitivity is in the range of several kilohertz per degree Kelvin.

16.8 Piezoelectric Temperature Sensors

Piezoelectric effect, in general, is a temperature-dependent phenomenon. Thus, a temperature sensor based on variability of oscillating frequency of a quartz crystal can be designed. Since the quartz is an anisotropic medium, the resonant frequency of a plate is highly dependent on the crystallographic orientation of the plate, the so-called angle of cut. Thus, by selecting a cut, negligibly small temperature sensitivity may be achieved (*AT*- and *BT*-cuts), or just the opposite – a cut with pronounced temperature dependence may be selected. Temperature dependence of the resonant frequency may be approximated by a third-order polynomial

$$\frac{\Delta f}{f_0} = a_0 + a_1 \Delta T + a_2 \Delta T^2 + a_3 \Delta T^3, \tag{16.53}$$

where ΔT and Δf are the temperature and frequency shifts, respectively, f_0 is the calibrating frequency, and a values are the coefficients. The first utilization of temperature dependence was made in 1962 by utilizing a nonrotated Y-cut crystal [26]. A very successful development of a linear temperature coefficient-cut (LC) was made by Hewlett–Packard [27]. The second- and third-order coefficients had been eliminated by selecting a doubly rotated *Y*-cut. The sensitivity (a_1) of the sensor is 35 ppm/°C, and the operating temperature range is from −80 to 230°C with calibration accuracy of 0.02°C. With the advent of microprocessors, linearity became a less important factor and more sensitive, yet somewhat nonlinear quartz temperature sensors had been developed by using a slightly singly rotated *Y*-cut

($Q = -4°$) with sensitivity of 90 ppm/°C [28] and by utilizing a tuning-fork resonators in flexural and torsional modes [29, 30].

It should be noted, however, that thermal coupling of the object of measurement with the oscillating plate is always difficult and; thus, all piezoelectric temperature sensors have relatively slow response when compared with thermistors and thermoelectrics.

References

1. Fraden J, Ferlito RK (2007) Ear temperature monitor and method of temperature measurement. US Patent 7,306,565, 11 Dec
2. Benedict RP (1984) Fundamentals of temperature, pressure, and flow measurements, 3rd edn. Wiley, New York
3. Callendar HL (1887) On the practical measurement of temperature. Philos Trans R Soc Lond 178:160
4. Sapoff M (1999) Thermistor thermometers. In: Webster JG (ed) The measurement, instrumentation and sensors handbook. CRC Press, Boca Raton, FL, pp 32.25–32.41
5. Fraden J (2000) A two-point calibration of negative temperature coefficient thermistors. Rev Sci Instrum 71(4):1901–1905
6. Steinhart JS, Hart SR (1968) Deep Sea Res 15:497
7. Mangum BW (1983) Rev Sci Instrum 54(12):1687
8. Sapoff M, Siwek WR, Johnson HC, Slepian J, Weber S (1982) In: Schooley JE (ed) Temperature. Its measurement and control in science and industry. American Institute of Physics, Washington, DC, vol 5, p 875
9. Villemant CM, Gaultier M (1971) Thermistor. US Patent 3,568,125, 2 Mar
10. Silver EH et al (2007) Method for making an epitaxial germanium temperature sensor. US Patent 7,232,487, 19 Jun
11. Tosaki H et al Thick film thermistor composition. US Patent 4,587,040, 6 May 1986
12. Keystone NTC PTC thermistors (1984) Catalogue © Keystone Carbon Company, St. Marys, PA
13. Caldwell FR (1962) Thermocouple materials. NBS monograph 40. National Bureau of Standards. March 1
14. Manual on the use of thermocouples in temperature measurement (1993) 4th edn. ASTM Manual Series: MNL: 12–93, ASTM, Philadelphia, PL
15. Sachse HB (1975) Semiconducting temperature sensors and their applications. Wiley-Interscience, New York
16. Timko MP (1976) A two terminal IC temperature transducer. IEEE J Solid-State Circuits SC-11:784–788
17. Wickersheim KA, Sun MH (1987) Fluoroptic thermometry. Med Electron 84–91
18. Fernicola VC et al (2000) Investigations on exponential lifetime measurements for fluorescence thermometry. Rev Sci Instrum 71(7):2938–2943
19. Schultheis L, Amstutz H, Kaufmann M (1988) Fiber-optic temperature sensing with ultrathin silicon etalons. Opt Lett 13(9):782–784
20. Wolthuis RA, Mitchell GL, Saaski E, Hartl JC, Afromowitz MA (1991) Development of medical pressure and temperature sensors employing optical spectral modulation. IEEE Trans Biomed Eng 38(10):974–981
21. Beheim G, Fritsch K, Azar MT (1990) A sputtered thin film fiber optic temperature sensor. Sensors 37–43
22. Hao T, Lui CC (1990) An optical fiber temperature sensor using a thermochromic solution. Sens Actuators A 24:213–216

23. Williams J (1990) Some techniques for direct digitization of transducer outputs, AN7. Linear technology application handbook
24. Venema A et al (1990) Acoustic-wave physical-electronic systems for sensors. In: Fortschritte der Acustik der 16. Deutsche Arbeitsgemeinschaft für Akustik, pp 1155–1158
25. Vellekoop MJ et al (1990) All-silicon plate wave oscillator system for sensor applications. In: Proc IEEE Ultrasonic Symp, New York
26. Smith WL, Spencer LJ (1963) Quartz crystal thermometer for measuring temperature deviation in the 10.3 to 10.6°C range. Rev Sci Instrum 268–270
27. Hammond DL, Benjaminson A (1962) Linear quartz thermometer. Instrum Control Syst 38:115
28. Ziegler H (1984) A low-cost digital sensor system. Sens Actuators 5:169–178
29. Ueda T, Kohsaka F, Iino T, Yamazaki D (1986) Temperature sensor utilizing quartz tuning fork resonator. In: Proc 40th annual frequency control symposium, Philadelphia, PA, pp 224–229
30. EerNisse EP, Wiggins RB (1986) A resonator temperature transducer with no activity dips. In: Proc 40th annual frequency control symposium, Philadelphia, PA, pp 216–223, 1986
31. Sola-Laguna LM et al (1998) Thick-film NTC thermistor series for sensor and temperature compensation application. In: IMAPS

Chapter 17
Chemical Sensors[1]

Of all smells, bread;
Of all tastes, salt.

–George Herbert, English poet

Sensors for measuring and detecting chemical substances are pervasively employed yet are, for the most part, unobtrusive. They are used to help run our cars more efficiently, track down criminals, and monitor our environment and health. Examples of uses include monitoring of oxygen in automobile exhaust systems, glucose levels in samples from diabetics, and carbon dioxide in the environment. In the laboratory, chemical detectors are the heart of key pieces of analytical equipment employed in the development of new chemicals and drugs and to monitor industrial processes. Progress has been impressive, and the literature is full of interesting developments. Recent developments include a broad spectrum of technologies, including improved screening systems for security applications [1] and miniaturization of systems once only used in laboratories [2]. Chemical sensors respond to stimuli produced by various chemicals or chemical reactions. These sensors are intended for identification and quantification of chemical species (including both liquid and gaseous phases).

In industry, chemical sensors are used for process and quality control during plastics manufacturing and in the production of foundry metals where the amount of diffused gasses affects metal characteristics such as brittleness. They are used for environmental monitoring of workers to control their exposure to dangers and limit health risks. Chemical sensors find many new applications as electronic noses. An electronic nose generally uses different types of sensor technologies [3] in order to mimic the olfaction capabilities of mammals [4]. In medicine, chemical sensors are used to determine patient health by monitoring oxygen and trace gas content in the lungs and in blood samples. These sensors are often used for breathalyzers to test

[1]This chapter is written in collaboration with Prof. Todd E. Mlsna (Mississippi State University, tmlsna@chemistry.msstate.edu) and Dr. Sanjay V. Patel (Seacoast Science, Inc., sanjay@seacoastscience.com).

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_17, © Springer Science+Business Media, LLC 2010

for blood alcohol levels and as indicators of the digestion problems of patients. In the military, chemical sensors are used to detect fuel dumps and to warn soldiers of the presence of airborne chemical warfare agents. Chemical sensors are used to detect trace contaminants in liquids, and, for example, they are used to search for and monitor ground water contamination near military, civilian, and industrial sites, where significant amounts of chemicals are stored, used, or dumped [5]. Combinations of liquid and gas sensors are used in experimental military applications to monitor compounds produced from refineries and nuclear plants to verify compliance with weapons treaties.

17.1 Overview

Traditionally, chemical sensing of unknown substances is done in an analytical laboratory with complex benchtop equipment including, for example, mass spectrometry, chromatography, nuclear magnetic resonance, X-ray, and infrared technology. These methods are very accurate, and it is possible to identify most classes of unknown chemicals with a high degree of confidence. However, the instruments are often expensive and require trained personnel to operate. Considerable efforts have been devoted to developing miniaturized low-cost sensing systems to address specific markets. Impressive advances have been made and many sensor systems are available at low cost, but these miniaturized systems traditionally have problems with sensitivity, selectivity, baseline stability, and reproducibility. Here, we provide an overview of chemical sensors and sensing systems for both the analytical laboratory and the miniaturized systems for specific applications.

17.2 History

The history of man benefiting from sensing chemicals is rich and colorful. The first examples involved clever use of the animal kingdom including the miner's canary [6] employed to monitor air quality (Fig. 17.1a). Early miners often worked in dangerous conditions without the benefit of modern ventilation systems. For hundreds of years, miners would work side by side with caged canaries to warn of dangerous environmental conditions. Canaries are more sensitive than man to low levels of methane, carbon monoxide, and diminished oxygen levels that can occur with a tunnel collapse or the release of pockets of trapped gases. As long as the canary lived the miner knew his air supply was safe. A dead canary alerted the miners of a potentially dangerous situation. In modern mines, most canaries have been replaced with various types of personal, portable, and fixed gas-monitoring equipment [7].

Since prehistoric times, canines were used for finding and tracking game. Today, trained dogs are used for finding explosives and drugs in airports and other public

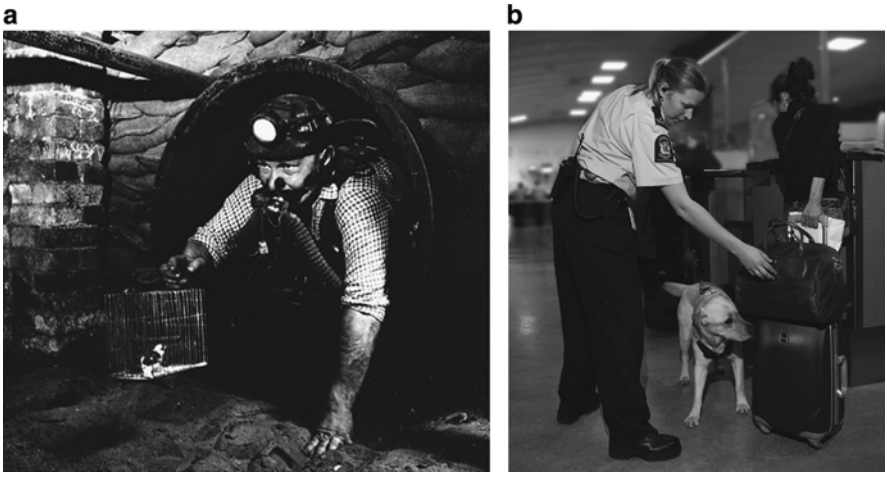

Fig. 17.1 Miner with caged canary (**a**). Customs Service drug detector dog (**b**)

places (Fig. 17.1b). A disproportionately large portion of the dog's brain is dedicated to smell when compared to the human brain. As a result dogs have demonstrated the ability to discriminate some odors at concentrations 8 orders of magnitude lower than man. Ever since dogs have been domesticated 30,000 years ago [8], man has relied on the dog nose. From these modest beginnings, chemical sensors have expanded beyond the animal kingdom, grown to become big business, and are now pervasively employed.

17.3 Chemical Sensor Characteristics

Most chemical sensors can be described using criteria and characteristics general to all sensors, such as stability, repeatability, linearity, hysteresis, saturation, response time, and span (*see* Chap. 2), but two characteristics are unique and meaningful as applied to chemical detection. Because chemical sensors are used for both identification and quantification, they need to be both selective and sensitive to a desired target species in a mixture of chemical species.

Selectivity describes the degree to which a sensor responds to only the desired target species, with little or no interference from nontarget species. Therefore, one of the most important functions in the evaluation of a chemical sensor's performance is the qualification of its selectivity.

Sensitivity describes the minimal concentrations and concentration changes (then referred to as resolution) that can be successfully and repeatedly sensed by a device. Figures 17.2 and 17.3 display typical sets of data used to establish a sensor's sensitivity and selectivity. Note that for the sensors described in the previous chapters, the term sensitivity is often used as a synonym of "slope"

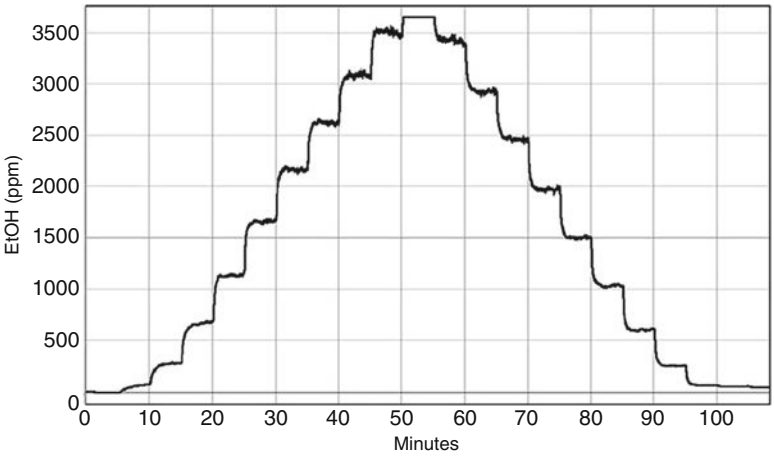

Fig. 17.2 Metal-oxide-semiconductor-based sensor response to increasing and decreasing concentrations of ethanol

when the transfer function of a sensor is linear. For the chemical sensors, sensitivity is the synonym of resolution.

17.4 Classes of Chemical Sensors

There is no universally accepted method to categorize the complete list of chemical detectors. For the purpose of this chapter, we have grouped them into two main sections, one being transduction method and the other being method of implementation. We have further divided the methods of transduction into three classes, including (1) sensors that measure electrical or electrochemical properties, (2) those that measure a change in a physical property, (3) and those that rely on optical absorption. An impressive range of sensor technologies have been developed to respond to different chemical, physical, and optical properties to aid in the detection of chemical analytes. Some of these technologies, for example microcantilevers, can be used to measure chemical and/or physical properties and thus are not easily classified.

17.4.1 Electrical and Electrochemical Transducers

Sensors that directly measure the electrical properties of a target analyte or the effect of the analyte on the electrical properties of another material are often the least expensive commercially available detectors. With these sensors, detection can be a reversible or a destructive irreversible process resulting in analyte decomposition.

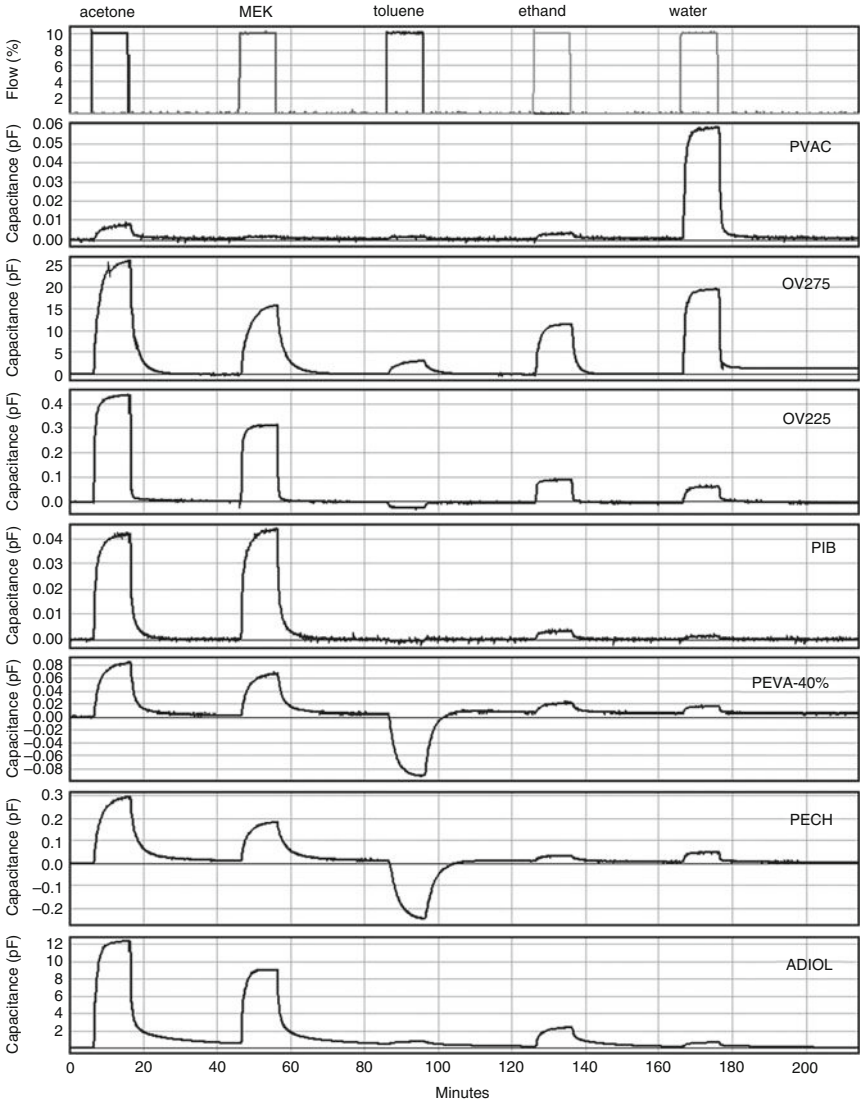

Fig. 17.3 Response of a capacitive VOC sensor array containing seven differently absorbing polymer-coated chemicapacitors to pulses of acetone, methyl ethyl ketone, toluene, ethanol, and water at 25°C

These devices and supporting electronics are often simple in design and resulting products can often be used in harsh applications. Sensors in this class include metal-oxide semiconductors, electrochemical sensors, potentiometric sensors, conductometric sensors, amperometric sensors, elastomeric chemiresistors, chemicapacitors, and chemFETs.

17.4.1.1 Metal-Oxide Semicondutor Devices

The most common type of metal-oxide-based sensor (MOS) translates changes in the concentration of a reactive species into changes in resistance. Development of these sensors began over 50 years ago when researchers discovered that the resistivity of a semiconductor changes with its chemical environment [9]. Germanium was used as an early model and clearly displayed measurable changes in resistance but suffered from problems with reproducibility for a range of reasons. Today, metal-oxide sensors are commercially available, very inexpensive, robust, and serve in a number of different applications.

A metal-oxide-based sensor is generally comprised of a semiconducting sensitive layer, an electrical connection to measure the resistance of that layer, and a heater to control the temperature of the device [9]. After a reactive molecule chemisorbs on the metal-oxide surface, a charge transfer takes place. When a metal-oxide crystal such as SnO_2 is heated at a certain high temperature in air, oxygen is adsorbed on the crystal surface, and a surface potential is formed that inhibits electron flow. When the surface is exposed to oxidizable gases such as hydrogen, methane, and carbon monoxide, the surface potential lowers and conductivity measurably increases [10]. As the concentration of the target chemical increases, so does the magnitude of the change in resistance.

The relationship between the film's electrical resistance and a given oxidizable gas's concentration is described by the following empirical equation:

$$R_s = A[C]^{-\alpha}, \tag{17.1}$$

where R_s is the sensor electrical resistance, A is a constant specific for a given film composition, C is the gas concentration, and α is the characteristic slope of R_s curve for that material and expected gas [11].

Metal-oxide devices change resistivity as the function of the presence of oxidizable gases, and as such they require additional electronic circuit to operate. A typical arrangement is to design the sensor as one leg in a common Wheatstone bridge circuit (*see* Sect. 5.11 in Chap. 5) so that the changing resistance can be detected as an unbalancing of the potential drops observed across the bridge circuit (Fig. 17.4a). The NTC thermistor[2] with a linearizing parallel resistor is required to adjust the bridge balance point according to the sensor's temperature.

Since the sensor behaves as a variable resistor whose value is controlled by gas species and gas concentration, the voltage drop across it is proportional to its resistance, and a plot of voltage drop versus gas concentration is recorded. The response signal from the sensors is linear when plotted on logarithmic charts (Fig. 17.4b). The slopes and offsets of the curves produced by different oxidizable gases allow them to be distinguished from each other and quantified within certain concentration ranges where the curves do not overlap [12]. Optionally, the rate of

[2]Resistive temperature sensor having negative temperature coefficient (NTC) – see Chap. 16.

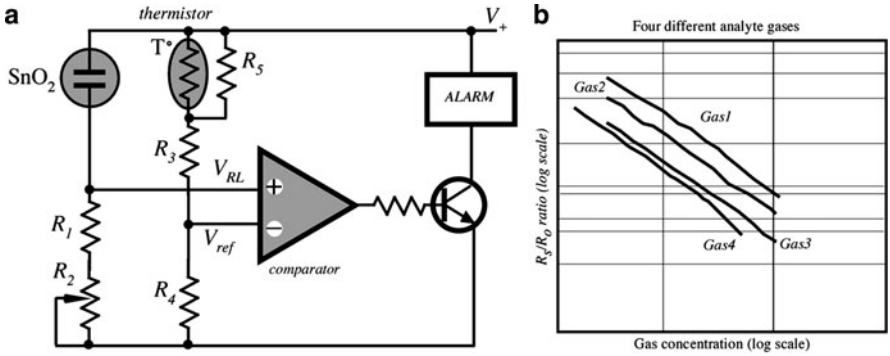

Fig. 17.4 SnO_2 Wheatstone bridge circuit (**a**) used for metal-oxide sensors and responses to different gases (**b**)

change of the conductivity may be used to differentiate gases and concentrations [13]. The bulk conductivity can drift for these devices, but the rate of change of that conductivity when driven by a pulsed input is more stable and reproducible.

These solid-state sensors have the advantage of being small, having low power consumption, low in cost, and can be easily batch fabricated. The control and measurement circuitry can be fabricated on the silicon microchip as well, providing opportunities to deploy sensor packages containing monolithically integrated arrays of sensing elements along with on-chip data acquisition and control systems. Several references to thin and thick film sensors on silicon devices have appeared based on a number of different materials for sensing a variety of gases [14, 15]. Tin oxide is the most prevalent pure film material studied [16, 17, 18, 19, 20]. In addition, Pt-doped [21, 22] and Pd-doped tin oxide films [23, 24] have been used to sense carbon monoxide, hydrogen, and hydrocarbons. Titania, in various forms and environments, has been used for sensing oxygen [25]. Rhodium-doped TiO_2 has been used to sense hydrogen [26]. Zinc oxide has been used to sense hydrogen, carbon monoxide and hydrocarbons [27]. The electrical properties of these materials change with the adsorption, absorption, desorption, rearrangement, and reaction of gases on the surface or in the bulk. Many of these materials have catalytic properties, and the adsorption and/or surface reactions of gases contribute to changes in electrical conductivity.

17.4.1.2 Electrochemical Sensors

The electrochemical sensors are commercially available and very versatile. Depending on the operating mode, they are divided into sensors that measure voltage (potentiometric), those that measure electric current (amperometric), and those that rely on the measurement of conductivity or resistivity (conductometric). In all these methods, special electrodes are used, where either a chemical reaction

takes place or the charge transport is modulated by the reaction. A fundamental rule of an electrochemical sensor is that it always requires a closed circuit, that is, an electric current (either dc, or ac) must be able to flow in order to make a measurement. Since electric current flow essentially requires a closed loop, the sensor needs at least two electrodes, one of which often is called a return electrode. It should be noted, however, that even if in a potentiometric sensor no flow of current is required for the voltage measurement, the loop still must be closed for measuring voltage.

The electrodes in these sensing systems are often made of catalytic metals such as platinum or palladium, or they can be carbon-coated metals. Electrodes are designed to have high-surface area to react with as much of the analyte as possible, producing the largest measurable signal. Electrodes can be treated (modified) to improve their reaction rates and extend their working life spans. The working electrode (WE) is where the targeted chemical reactions take place (Fig. 17.5). The electrical signal is measured with respect to a counter or auxiliary electrode (AE) that is not intended to be catalytic, and in the case of three-electrode systems, a third reference electrode (RE) is employed to measure and correct for electrochemical potentials generated by each electrode and the electrolyte. The third electrode improves operation by correcting for error introduced by a polarization of the working electrode. Newer electrochemical sensors employ thick-film, screen-printed electrode sets to make manufacturing simpler and more robust.

The electrolyte is a medium that carries charges using ions instead of electrons. This directly limits the reactions that can take place and is the first stage of lending selectivity to the electrochemical sensor. The sensor formed by this collection of

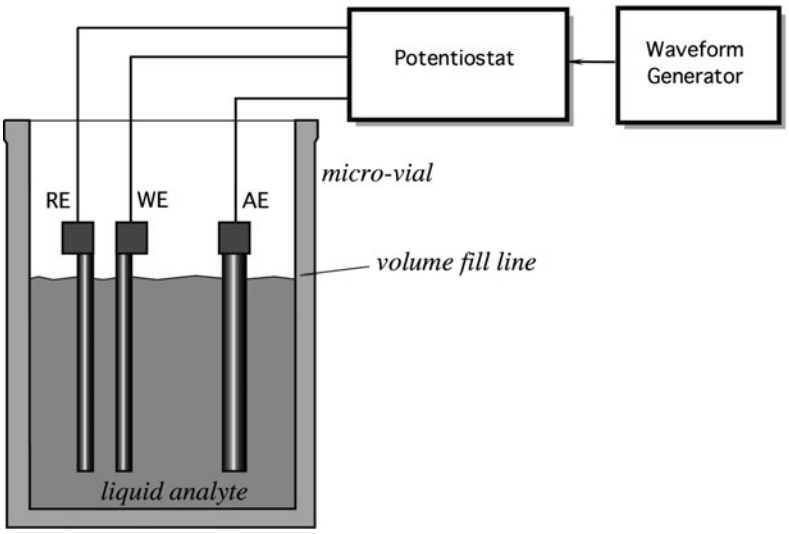

Fig. 17.5 Electrochemical-sensor electrode set

electrodes and electrolytes is called an electrochemical cell and may be operated in several ways depending upon the electrical characteristic being observed (resistance, potential, current, capacitance, etc.). The more comprehensive measurements are captured using various forms of voltammetry discussed later in this chapter.

A simple liquid electrochemical sensor (cell) uses two electrodes immersed in an electrolyte solution. Gas analytes, such as CO, react at the working electrode and produces CO_2 and free electrons. Charges and charged species migrate to the other (counter) electrode where water is produced if oxygen is present. The reaction converts CO to CO_2 (17.2). If the electrodes are connected in series to a resistor and the potential drop across the resistor is measured, it will be proportional to the current flowing making it a function of analyte gas present.

$$\mathrm{CO} + \mathrm{H_2O} \leftrightarrow \mathrm{CO_2} + 2\mathrm{H}^+ + 2\mathrm{e}^- \tag{17.2}$$

17.4.1.3 Potentiometric Sensors

These sensors use the effect of the concentration on the equilibrium of the redox reactions occurring at the electrode–electrolyte interface in an electrochemical cell. An electrical potential may develop at this interface due to the redox reaction that takes place at the electrode surface, where Ox denotes the oxidant, Ze^- is the number of electrons involved in the redox reaction, and Red the reduced product [28]

$$Ox + Ze^- = Red. \tag{17.3}$$

This reaction occurs at one of the electrodes (cathodic reaction in this case) and is called a half-cell reaction. Under thermodynamical quasiequilibrium conditions, the Nernst equation is applicable and can be expressed as

$$E = E_0 + \frac{RT}{nF} \ln \frac{C_{\mathrm{O}}^*}{C_R^*}, \tag{17.4}$$

where C_{O}^* and C_R^* are the concentrations of Ox and Red, respectively, n is the number of electrons transferred, F is the Faraday constant, R is the gas constant, T is the absolute temperature, and E_0 is the electrode potential at a standard state. In a potentiometric sensor, two half-cell reactions will take place simultaneously at each electrode. However, only one of the reactions should involve the sensing species of interest, while the other half-cell reaction is preferably reversible, noninterferring, and known.

The measurement of the cell potential of a potentiometric sensor should be made under zero-current or quasiequilibrium conditions; thus, a very high input impedance amplifier (which is called an electrometer) is generally required. There are two

types of electrochemical interfaces from the viewpoint of the charge transfer: ideally polarized (purely capacitive) and nonpolarized. Some metals (e.g. Hg, Au, Pt) in contact with solutions containing only inert electrolyte (e.g. H_2SO_4) approach the behavior of the ideally polarized interface. Nevertheless, even in those cases, a finite charge-transfer resistance exists at such an interface and excess charge leaks across with the time constant given by the product of the double-layer capacitance and the charge-transfer resistance ($\tau = R_{ct} C_{dl}$).

An ion-selective membrane is the key component of all potentiometric ion sensors. It establishes the reference with which the sensor responds to the ion of interest in the presence of various other ionic components in the sample. An ion-selective membrane forms a nonpolarized interface with the solution. A well-behaved membrane, i.e. one that is stable, reproducible, immune to adsorption and stirring effects, and also selective, has both high absolute and relative exchange-current density.

17.4.1.4 Conductometric Sensors

An electrochemical conductivity sensor measures the change in conductivity of the electrolyte in an electrochemical cell. An electrochemical sensor may involve a capacitive impedance resulting from the polarization of the electrodes and faradic or charge transfer process.

In a homogeneous electrolytic solution, the conductance of the electrolyte G (per Ω), is inversely proportional to L, which is the segment of the solution along the electrical field and directly proportional to A, which is the cross-sectional area perpendicular to the electric field [29]

$$G = \frac{\rho A}{L}, \tag{17.5}$$

where ρ (per Ω/cm) is the specific conductivity of the electrolyte and is related quantitatively to the concentration and the magnitude of the charges of the ionic species. According to Kohlrausch [30], the equivalent conductance of the solution at any concentration, C in mol/l or any convenient units, is given by

$$\Lambda = \Lambda_0 \beta \sqrt{C}, \tag{17.6}$$

where β is a characteristic of the electrolyte, and Λ_0 is the equivalent conductance of the electrolyte at an infinite dilution.

Measurement techniques of electrolytic conductance by an electrochemical conductivity sensor have remained basically the same over the years. Usually, a Wheatstone bridge (similar to Fig. 17.4) is used with the electrochemical cell (the sensor) forming one of the resistance arms of the bridge. However, unlike the measurement of the conductivity of a solid, the conductivity measurement of an

electrolyte is often complicated by the polarization of the electrodes at the operating voltage. A faradic or charge transfer process occurs at the electrode surfaces. Therefore, a conductivity sensor should be operated at a voltage where no faradic process could occur. Another important consideration is the formation of a double layer adjacent to each of the electrodes when a potential is imposed on the cell. This is described by the so-called Warburg impedance. Hence, even in the absence of the faradic process, it is essential to take into consideration the effect of the double layers during measurement of the conductance. The effect of the faradic process can be minimized by maintaining the high cell constant L/A of the sensor so that the cell resistance lies in the region between 1 and 50 kΩ. This implies using a small electrode surface area and large interelectrode distance. This, however, reduces the sensitivity of the Wheatstone bridge. Often the solution is in use of a multiple electrode configuration. Both effects of the double layers and the faradic process can be minimized by using a high frequency low-amplitude alternating current. Another technique is to balance both the capacitance and the resistance of the cell by connecting a variable capacitor in parallel to the resistance of the bridge area adjacent to the cell.

17.4.1.5 Amperometric Sensors

An example of an amperometric chemical sensor is a Clark oxygen sensor that was proposed in 1956 [31]. The operating principle of the electrode is based on the use of electrolyte solution contained within the electrode assembly to transport oxygen from an oxygen-permeable membrane to the metal cathode. The cathode current arises from a two-step oxygen-reduction process that may be represented as

$$\begin{aligned} O_2 + 2H_2O + 2e^- &\rightarrow H_2O_2 + 2OH^- \\ H_2O_2 + 2e^- &\rightarrow 2OH^-. \end{aligned} \tag{17.7}$$

Figure 17.6a shows the membrane that is stretched across the electrode tip, allowing oxygen to diffuse through a thin electrolyte layer to the cathode. Both anode and cathode are contained within the sensor assembly, and no electrical contact is made with the outside sample. A first-order diffusion model of the Clark electrode is illustrated in Fig. 17.6b [32]. The membrane–electrolyte–electrode system is considered to act as a one-dimensional diffusion system with the partial pressure at the membrane surface equal to the equilibrium partial pressure p_0 and that at the cathode equal to zero. It can be shown that the equilibrium steady-state electrode current is given by

$$I \approx \frac{4Fa_m D_m p_0}{x_m}, \tag{17.8}$$

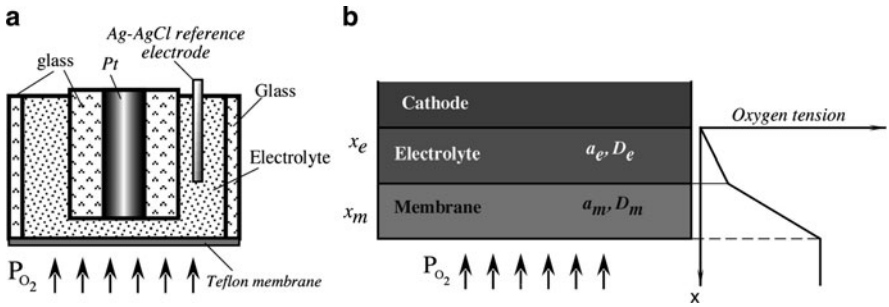

Fig. 17.6 Clark electrode (**a**); and the first-order one-dimensional model (**b**) of the oxygen tension distribution throughout the system (adapted from Clark [32])

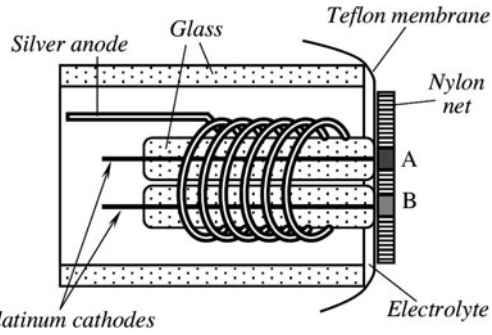

Fig. 17.7 Simplified schematic of an amperometric Clark oxygen sensor adapted for detecting glucose

where A is the electrode area, α_m is the solubility of oxygen in the membrane, F is the Faraday's constant, D_m is the diffusion constant, and x_m is the thickness of the membrane. It should be noted that the current is independent on the electrolyte thickness and diffusion properties. A Teflon® membrane is used as an oxygen-permeable film. We may define the sensor's sensitivity as a ratio of the current to the oxygen partial pressure

$$S = \frac{I}{p_0}. \tag{17.9}$$

For example, if the membrane is 25 μm thick and the cathode area is 2×10^{-6} cm^2, then the sensitivity is approximately 10^{-12} A/mmHg.

An enzymatic type amperometric sensor can be built with a sensor capable of measuring the relative oxygen deficiency caused by the enzymatic reaction by using two Clark oxygen electrodes. The operating principle of the sensor is shown in Fig. 17.7. The sensor consists of two identical oxygen electrodes, where one (A) is coated with an active enzyme layer, while the other (B) with an inactive enzyme layer. An example of the application is a glucose sensor, where inactivation can be

carried out either chemically, or by radiation, or thermally. The sensor is encapsulated into a plastic carrier with glass coaxial tubes supporting two Pt cathodes and one Ag anode. In the absence of the enzyme reaction, the flux of oxygen to these electrodes, and therefore the diffusion-limiting currents, are approximately equal to one another. When glucose is present in the solution and the enzymatic reaction takes place, the amount of oxygen reaching the surface of the active electrode is reduced by the amount consumed by the enzymatic reaction, which results in a current imbalance.

17.4.2 Elastomer Chemiresistors

Elastomer chemiresistors or polymer conductive composites (also polymer conductors or simply "PCs") are polymer films that adsorb chemical species and swell, increasing resistance as a physical response to the presence of a chemical species. These can be used as chemical detectors but do not truly employ a chemical reaction. The polymers are designed and/or treated to attract subsets of chemicals providing a degree of speciation or selectivity [33]. The PC sensors can respond to the presence of simple hydrocarbons like isopropyl alcohol in only a couple of seconds, while more complex oils may take 10–15 s. The PC element is not expected to be tolerant of corrosives, but barring exposure to such should have a life span of months in normal operation. The PC measurement strategy uses several differently treated PC elements to produce an array, and then samples the array to produce a signature. Unlike metal-oxide-based sensors, the PCs do not require the high, controlled operating temperatures and therefore consume significantly less power.

To detect the presence of a liquid, a sensor usually must be specific to that particular agent at a certain concentration. That is, it should be selective to the liquid's physical and/or chemical properties. An example of such a sensor is a resistive detector of hydrocarbon fuel leaks. A detector is made of silicone (a nonpolar polymer) and carbon black composite. The polymer matrix serves as the sensing element and the conductive filler is used to achieve a relatively low-volume resistivity, on the order of 10 Ω cm in the initial stand-by state. The composition is selectively sensitive to the presence of a solvent with a large solvent–polymer interaction coefficient [34], and the resistance can be modified by varying the conductive particle to polymer ratio. Since the sensor is not susceptible to polar solvents such as water or alcohol, it is compatible with the underground environment. The sensor is fabricated in the form of a thin film with a very large surface/thickness ratio. Whenever the solvent is applied to the thin-film sensor, the polymer matrix swells resulting in the separation between conductive particles. This causes a conversion of the composite film from being more conductive to less conductive with a resistivity on the order $10^9\Omega$ cm, or even higher. The response time for a thin-film sensor is less than 1 s. The sensor returns to its normally conductive state when it is no longer in contact with the hydrocarbon fuel, making the device reusable.

17.4.2.1 Chemicapacitive Sensors

A "chemicapacitive" sensor (or "chemicapacitor") is a capacitor that has a selectively absorbing material such as a polymer or other insulator, as a dielectric. When a chemical absorbs into the dielectric, its permittivity is altered and correspondingly the capacitance of the sensor changes[35]. The most common type of commercially available chemicapacitor consists of water-sensitive polymers and is used for humidity sensing (*see* Sect. 13.2 in Chap. 13). However, chemicapacitors are not limited to polymer dielectrics. Other materials have been used to broaden the range of detectable chemicals, sol–gel chemicapacitors, for example, can detect carbon dioxide [36] – although such materials often have to be heated to achieve optimal performance. More recently, polymers have been used to make low-power sensors for volatile organic compounds (VOCs) [37].

Chemicapacitors can be constructed using conventional thin film techniques, where conductive electrodes are arranged in either a parallel or interdigital layout. Typically, interdigitized electrodes consist of a single layer of metal deposited on a substrate to form two meshed combs. The polymer or other materials are deposited on top of the combs. Parallel-plate sensors typically consist of a layer of metal deposited on a substrate, followed by a layer of insulator and finally a second, porous layer of metal on top of the insulator [38].

A robust MEMS-sensor based on micromachined capacitors has been developed and commercialized [39]. An example geometry is the square-shaped parallel-plate capacitor seen in Fig. 17.8. It has approximately 285 μm on a side, with a 0.75 μm vertical gap between the plates (Fig. 17.9). The top plate is perforated forming a waffle pattern, with silicon beams of 2.5 μm and holes of 5 μm. The 16 larger squares are the support posts, which together with the outer edge of the square (also perforated) keep the top plate from flexing. The structures are made from conductive polycrystalline silicon, deposited on an insulating silicon nitride layer using commercially available semiconductor manufacturing methods [40]. The chips have a standard wafer thickness of approximately 300 μm.

These types of sensor chips can be made with varying geometries and varying numbers of sensors and each sensor can receive a different analyte-sensitive coatings.

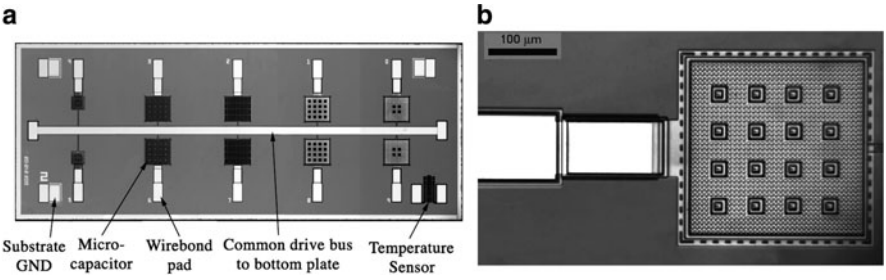

Fig 17.8 MEMS chip (2 × 5 mm) containing a variety of parallel-plate capacitor designs (**a**). Close-up top-view of a parallel-plate capacitor (**b**)

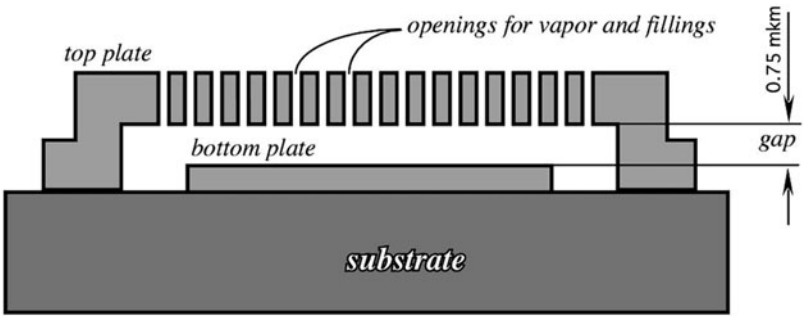

Fig. 17.9 Cross-section diagram of the parallel-plate capacitor showing the 0.75 μm gap

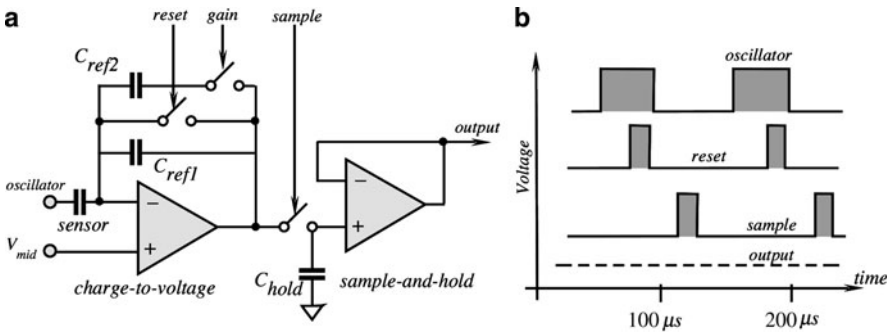

Fig. 17.10 Capacitance measurement circuit (**a**) and timing diagrams (**b**)

Each capacitor is filled with a polymer using an ink-jet [37]. The interaction between target analyte and polymer modifies the dielectric properties of the polymer resulting in a change in capacitance. Any capacitance measuring circuit can be used to measure these types of devices. These MEMS detector arrays operate well in ambient air at ambient pressures and temperatures, thus requiring no special compressed carrier gas and allowing for systems with decreased size and increased portability. They are now used commercially as the detector for a gas chromatograph suitable for training students in academic laboratories.[3]

To measure the capacitance, a circuit applies a square wave to the bottom plate. A charge/discharge readout circuit [41, 42] shown in Fig. 17.10 measures the capacitance of each sensor array using an oscillating charge/discharge drive voltage and producing the corresponding output voltage, V_{out}:

$$C_{Sensor} = \frac{(V_{out} - V_{mid})}{\Delta V_{osc}} \left(\Sigma C_{ref}\right), \tag{17.10}$$

[3]Vernier Mini Gas Chromatograph (www.vernier.com/probes/gc-mini.html).

where V_{mid} is a virtual ground voltage or reference, ΔV_{osc} the amplitude of the oscillator drive voltage, C_{Sensor} the capacitance of the capacitive sensor, and C_{ref} is the reference capacitance. In this example, the value of the reference capacitance is either 1 or 0.5 pF and is determined by the position of the gain switch. In the circuit, the reference capacitors are charged as the sensing capacitor discharges.

17.4.2.2 ChemFET

A ChemFET is a chemical field effect transistor that includes a gas-selective coating or series of coatings between its transistor gate and the analyte (Fig. 17.11). This chemical element gives the device a control input that modifies the source–drain conduction in relationship with selected chemical species. Different materials applied to the gate react with different chemical species (gases or liquids) and provide differentiation of species. ChemFETs can be used for detecting H_2 in air, O_2 in blood, some military nerve gases, NH_3, CO_2, and explosive gases [43]. A typical packaging of the gas sensor is shown in Fig. 7.12.

As in a conventional FET, the ChemFET is constructed using thin film techniques and commonly employs a p-type silicon body with two n-type silicon diffusion regions (source and drain). This three-part system is covered with an insulating layer, e.g. silicon dioxide, separating a final top metal gate electrode above and between the source and the drain. Operation involves applying a voltage, positive

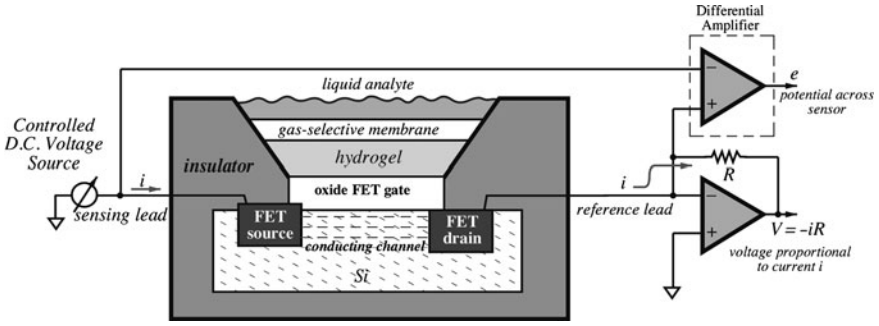

Fig. 17.11 Liquid ChemFET construction and electrical connection

Fig. 7.12 Gas sensor packaging

with respect to the silicon to the gate electrode. Electrons are attracted to the surface of the semiconductor forming a conducting channel between the source and the drain n-regions [44]. In fact, a ChemFET is a chemically controlled conductor (resistor). Conductance of ChemFET is measured by a differential amplifier and is represented by the output voltage e. To compute conductance, the current in the circuit is measured by the i/V converter with a reference resistor R.

Hydrogen gas-sensing chemFETs use a palladium/nickel (Pd/Ni) film as their gates [45]. The improved, more stable, chemFETs that are used for liquid sensing employ a silver/sliver chloride hydrogel (Ag/AgCl) bridge between the silicon dioxide (SiO_2) gate and a selective membrane that separates the gate from the analyte (Fig. 17.11). The selective membrane is commonly polyvinyl chloride (PVC), polyurethane, silicone rubber, or polystyrene.

For an ion-selective ChemFET (ISFET), the gate is replaced by or coated with a chemical-selective electrolyte or other semiconductor material [46]. If the ion-sensitive material is ion-penetrable, then the device is called a MEMFET; and if the membrane is ion-impenetrable, it is called a SURFET, since a surface potential is established by the ions. The chemical-selective gate material alters the potential at which the device begins to conduct and thus indicates the presence of specific chemical species. The devices are inherently small and low in power consumption. The gate coatings for the ChemFET can be enzyme membranes (ENFET) or ion-selective membranes. Ion-selective membranes produce a chemical sensor, while enzyme membranes can produce a biochemical sensor. The enzyme membrane is made from polyaniline and is itself created using a voltammetric electrochemical process to produce this organic semiconductor [47].

17.4.3 Photoionization Detector

A photoionization detector (PID) typically uses high-energy ultraviolet (UV) light to break molecules into positively charged ions. The molecules absorb the light energy resulting in temporary loss of electrons and the formation of positively charged ions. The molecules produce an electric current, which is measured with an electrometer.[4] Equation (17.11) shows a molecular species, R, being ionized by incident UV radiation, to an ion, R^+ and electron.

$$R + h\nu \rightarrow R^+ + e^- \tag{17.11}$$

The UV lamp is the heart of the detector and improvements in the UV lamp designs have led to significant reduction in cost and longer life expectancies. The wavelength of the UV light depends on the type of gas in the lamp employed. A popular

[4]An electrometer is an instrument for measuring very small electric charges, currents, or electrical potential differences. It is characterized by very low leakage currents, down to 1 fA.

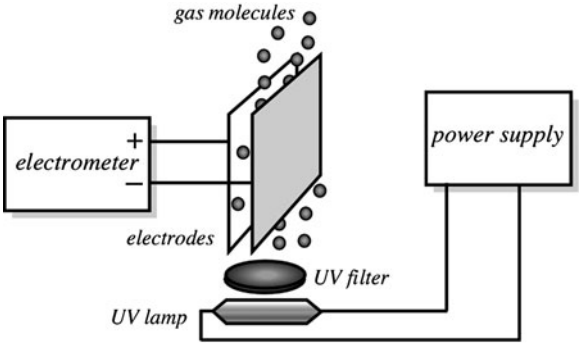

Fig. 17.13 Concept of PID detector

choice is krypton, which will emit light at with energies[5] of 10.0 and 10.6 eV. Xenon and argon lamps are also used occasionally.

When gas molecules pass by the UV lamp, they become ionized (Fig. 17.13). The free electrons are collected at a pair of closely placed electrode plates. These electrodes generate a signal in response to small changes in the electric field. The magnitude of the current flow is directly proportional to the gas concentration.

Each chemical has an ionization potential (IP), and gases with IP values below the rated eV output of the lamp will be ionized and thereby detected. For example, organic aromatic compounds and amines can be ionized by the 9.5 eV lamps, many aliphatic organic compounds require a 10.6 eV lamp and compounds such as acetylene, formaldehyde, and methanol require an 11.7 eV lamp. Each lamp can ionize gases with ionization potentials below its eV rating but will not ionize gases with higher ionization potentials. Typically, the portable units are equipped with the 10.6 eV lamp because of its ability to ionize most VOCs. Isobutylene is often used to calibrate these units. The output of the PID sensor is typically linear below 200 ppm and will become saturated above 2,000 ppm.

17.4.4 Physical Transducers

Several types of chemical sensors rely on measurement of a physical property of an analyte or the affect of the analyte's interaction with another material for detection. Usually, no chemical reaction occurs at the sensing element. These sensor technologies can be reversible or destructive. Reversible technologies include those that require absorption of the analyte into a substrate sitting on a sensitive microbalance that can respond to changes in mass. These sensors include surface acoustic wave

[5]The electron volt (eV) is a unit of energy. By definition, it is equal to the amount of kinetic energy gained by a single unbound electron when it accelerates through an electrostatic potential difference of 1 V. One eV is equal to $1.60217653 \times 10^{-19}$ J.

(SAW) devices, quartz crystal microbalances (QCMs), and microcantilevers. Destructive sensors may directly measure the molecular mass of an analyte, as in ion mobility spectrometry (IMS), or the quantity of heat released during complete oxidation, as in thermal or calorimetric sensors.

17.4.4.1 Acoustic Wave Devices

Acoustic wave devices can be used to make chemical sensors that detect very small mass change from adsorbed chemical molecules altering mechanical properties of a system and are referred to as mass, gravimetric, or microbalance sensors. These devices are generally constructed from piezoelectric crystals or materials that can be oscillated at high frequencies (from kHz to GHz). In various types of these devices, acoustic waves are generated by an oscillator circuit, which allows the crystal to resonate. The resonant frequency of the sensor changes when the crystal is perturbed, typically the frequency decreases when the mass of the device increases during sorption [48]. The shift in the resonant frequency of a piezoelectric crystal is proportional to the additional mass that is deposited on the crystal surface. Depending upon how the circuit is constructed, a piezoelectric quartz oscillator resonates with a frequency that is called either a series (f_r) or a parallel (f_{ar}) resonant (*see* Fig. 7.42b). Either frequency is a function of the crystal mass and shape. For example, in one type of sensing structure, which in a simplified manner may be described as an oscillating plate whose natural frequency depends on its mass, the mass change and frequency are related by

$$\frac{\Delta f}{f_0} = S_m \Delta m \tag{17.12}$$

where f_0 is the unloaded natural oscillating frequency, Δf is the frequency shift: $\Delta f = f_{\text{loaded}} - f_0$, Δm is the added mass per unit area, and S_m is called the sensitivity factor. The numerical value of S_m depends upon the design, material, and operating frequency (wavelength) of the acoustic sensor [48]. Since frequency and time are the easiest variables to measure by the electronic circuits, the entire sensor's accuracy is determined virtually by the ability to assure that coefficient S_m is known and does not change during the measurement. Figure 17.14 shows an example of this type of a sensor.

An electronic circuit measures the frequency shift, which can be related to a chemical concentration in the sampled gas. The absolute accuracy of the method depends on such factors as the mechanical clamping of the crystal, temperature, etc., therefore, calibration is usually required.

Generally speaking, there are four types of acoustic sensors widely used in chemical sensing research and products. These include QCMs, SAW, acoustic plate mode (APM), and flexural-plate-wave (FPW) devices made from thin membranes. There are also several variations on these types of devices, which have been developed or adapted for specific uses. These variations include use of different

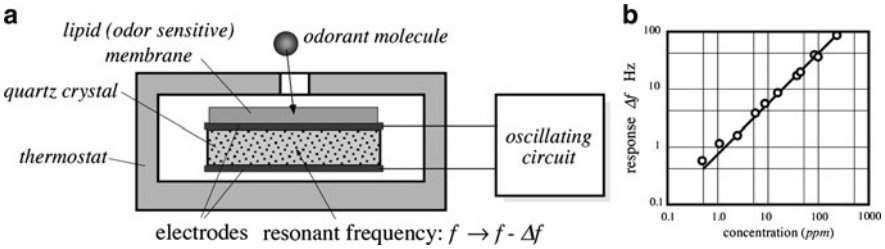

Fig. 17.14 Microbalance vapor sensor (**a**) and its transfer function (**b**) for amylacetate gas

modes of oscillation, resonance, and materials. Unlike a QCM, which operates at its resonant frequency, SAW, APM, and FPW devices are typically called "delay-line" devices. The delay refers to the time it takes for an applied electrical signal at one end (transmitter) of the device to propagate through the material (the acoustic wave) and to be measured at the opposing end (receiver).

Unlike the chemiresistors or chemicapacitors described in earlier sections, the gravimetric transducers do not need to directly measure the properties of a sensitive layer; rather they indirectly measure the interaction of the layer with the environment. Generally speaking, all of the oscillating sensors are extremely sensitive. For instance, a typical sensitivity is on the range of 5 MHz cm^2/kg, which means that 1 Hz in frequency shift corresponds to about 17 ng/cm^2 added mass. The dynamic range is quite broad: up to 20 μg/cm^2. To improve selectivity, devices may be coated with a chemical layer specific for the material of interest. Typical designs of the acoustic sensors that can be adapted for measuring mass are covered in Sect. 12.7. Here, we briefly describe the gravimetric sensors that are adapted for sensing gas concentrations (e.g. Fig. 17.15).

One type of a gravimetric detector is a SAW sensor. SAW devices use propagating mechanical waves along a solid surface, which is in contact with a medium of lower density, such as air [49]. These waves are sometimes called Rayleigh waves after the man who predicted them in 1885. As with the other delay-line devices, the SAW sensor is a transmission line with three essential components: the piezoelectric transmitter, the transmission line, typically with a chemically selective layer, and the piezoelectric receiver. An electrical oscillator causes the electrodes of the transmitter to flex the substrate, thus producing a mechanical or acoustic wave. The wave propagates along the transmission surface toward the receiver. The substrate may be fabricated of, e.g. $LiNbO_3$, a material with a high piezoelectric coefficient [50]. However, the transmission line does not have to be piezoelectric, which opens several possibilities of designing the sensor of different materials, like silicon. The transmission surface interacts with the sample according to the selectivity of the coating, thus modulating the propagating waves. The waves are received at the other end and subsequently converted back to an electric form. Often, there is another reference sensor whose signal is subtracted from the test sensor's output.

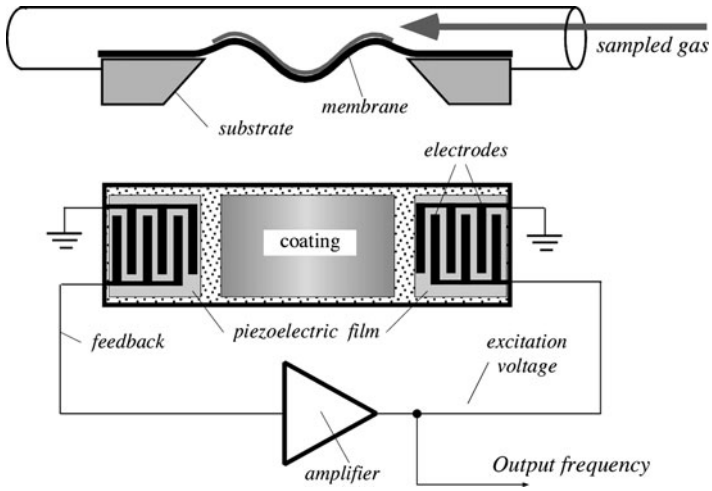

Fig. 17.15 Flexural plate SAW gas sensor (deflection of the membrane is exaggerated for clarity)

Another type of a gravimetric sensor is shown in Fig. 17.15. The sensor is designed in the form of a flexural thin silicon plate with two pairs of the interdigitized electrodes deposited by use of the sputtering technology. A thin piezoelectric ZnO thin film is deposited beneath the electrodes, so that the plate can be mechanically excited by the external electronic circuit. The piezoelectric film is needed to give silicon substrate piezoelectric properties. The top surface of the sensing plate is coated with a thin layer of a chemically selective material (or glue, if the sensor is intended to detect air pollutants). The entire sensor is positioned inside a tube where the sampled gas is blown through. The left and right pairs of the electrodes are connected to the oscillating circuit whose frequency f_0 is determined by the natural mechanical frequency of the sensor's plate.

The circuit contains an amplifier whose output drives the excitation electrode. Thanks to a piezoelectric effect, this causes flexing the membrane and propagation of the deflection wave from right to left. The wave velocity is determined by the state of the membrane and its coating. Change in the mechanical properties of the coating depends on its interaction with the sampled gas. Thus, the left electrodes will detect piezoelectric response either sooner or later, depending how fast the wave goes through the membrane. The received signal is applied to the amplifier's input as a feedback voltage and causes the circuit to oscillate. The output frequency is a measure of the sampled gas concentration. The reference frequency is usually determined before sampling the gas.

A theoretical sensitivity of the flexural plate sensor is given by $S_m = ½\rho d$, where ρ is the average density of the plate and d is its thickness [51]. At an operating frequency of 2.6 MHz, the sensor has sensitivity on the order of -900 cm^2/g. So, for example, if the sensor having the sensing area of 0.2 cm^2 captures 10 ng (10^{-8} g) of

material, the oscillating frequency is shifted by $\Delta f = -900 \times 2.6 \times 10^6 \times 10^{-8}/0.2 = -117$ Hz. Acoustic sensors are quite versatile and can be adapted for measuring variety of chemical compounds. The key to their efficiency is the selection of coating. Table 17.1 gives examples of various coatings for acoustic sensors.

17.4.4.2 Microcantilever

Microcantilevers are devices shaped like miniature diving boards that are typically micromachined from silicon or other materials. Originally used in various types of surface probe microscopies (SPM) [53], they have since been adapted to detect chemicals [54, 55] and biological materials [56, 57]. As with chemiresistors and acoustic devices, a chemically sensitive sorbent coating can be added to the cantilever to enhance its sensitivity and selectivity to certain chemicals. Cantilevers have been shown to detect a wide range of chemical analytes from fixed gases such as hydrogen [42] through common VOCs [58] to explosives [59].

The length of these cantilevers is often in the range of 100–200 μm, with a thickness range from 0.3 to 1 μm. Since 1994, these devices have been developed for detecting a variety of chemicals by monitoring either the bending or the frequency shift of an oscillating microcantilever [60, 61]. The key to the sensitivity of these very thin devices is the high surface-to-volume ratio. This amplifies the effect of surface stresses due to interactions with chemicals.

When oscillated at its harmonic frequency, a cantilever can be used to detect sorbed mass, much like the acoustic devices described earlier. The resonance frequencies decrease due to the adsorbed mass, and the more the mass absorbed, the greater the shift in frequency. Alternatively, the bending of a cantilever can be used to measure the change in surface stresses on the cantilever beam when a target chemical is preferentially absorbed to one of the surfaces of the cantilever by placing a selectively absorbent chemical coating on that surface (Fig. 17.16).

Table 17.1 SAW chemical sensor coatings and materials (after Nieuwenhuizen et al. [52])

Compound	Chemical coating	SAW substrate
Organic vapor	Polymer film	Quartz
SO_2	TEA[a]	Lithium niobate
H_2	Pd	Lithium niobate, silicon
NH_3	Pt	Quartz
H_2S	WO_3	Lithium niobate
Water vapor	Hygroscopic	Lithium niobate
NO_2	PC[b]	Lithium niobate, quartz
NO_2, NH_3, NH_3, SO_2, CH_4	PC[b]	Lithium niobate
Vapor explosives, drugs	Polymer	Quartz
SO_2, methane	C[c]	Lithium niobate

[a]TEA is triethanolamine
[b]PC is phthalocyanine
[c]No chemical coating used. Detection is based on changes in thermal conductivity produced by the gas

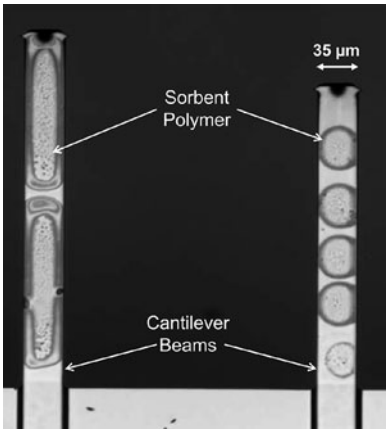

Fig. 17.16 Standard optical microcantilevers used for surface probe microscopy that have been modified with an absorbent coating. The cantilever beam on the right has five drops of a sorbent polymer coating, and the beam on the left has a continuous sorbent polymer coating covering the entire length

Because surface stress is being monitored, diffusion into the coating is not necessary; therefore, monolayer coatings are ideally suited for these devices.

The bending of the microcantilever is not caused by the weight of the absorbed chemical but by the absorption-induced surface stresses due to changes in surface free energy. The cantilever will bend if the surface free-energy density change is comparable with the cantilevers spring constant. When chemicals come in contact with a coated cantilever, electrostatic repulsions, swelling, or other effects result in changes in surface stress, which ultimately result in measurable cantilever bending.

Cantilevers can be measured in many ways. Originally, systems were based on optical-based (laser) detection that was developed for SPM (Fig. 17.17). Newer research has implemented thermal [62], capacitive [41], and piezoresistive [63] measurement techniques, thereby removing the need for lasers and associated optics, leaving a simpler measurement circuit.

17.4.4.3 Ion Mobility Spectrometry

IMS is a technique that detects and differentiates chemicals based on differential migration of ions under the influence of an electric field [64]. In IMS, gas-phase species are required to be ionized, for example, by high energy electrons from a radioactive ^{63}Ni source. Ions travel in a gas stream through an electric deflection field that spatially separates the ions with respect to their ion mobility at atmospheric pressure. Different ion species with different characteristics (mass, charge, and size) have different drift velocities, as in (17.13), where K is an ions mobility and v_d is its drift velocity (*see* Fig. 17.18). Ideally, individual ion beams develop, which are spatially separated at the bottom electrode setup.

$$v_d = KE \tag{17.13}$$

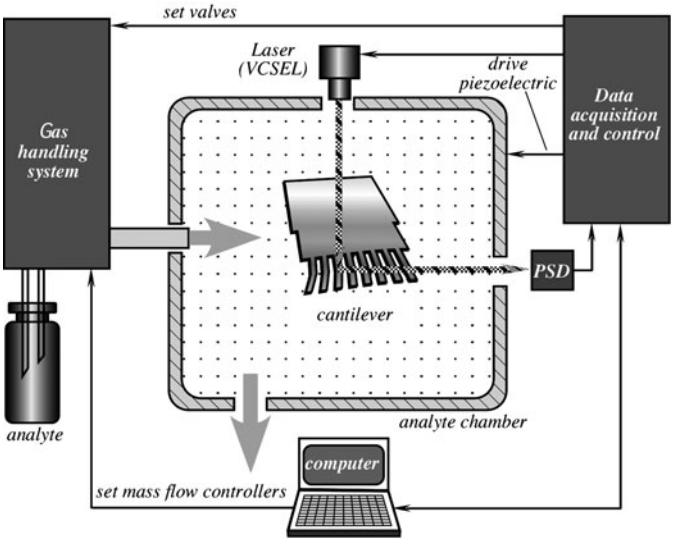

Fig. 17.17 Concept of the measurement setup for optical cantilevers (adapted from Battiston et al. [54])

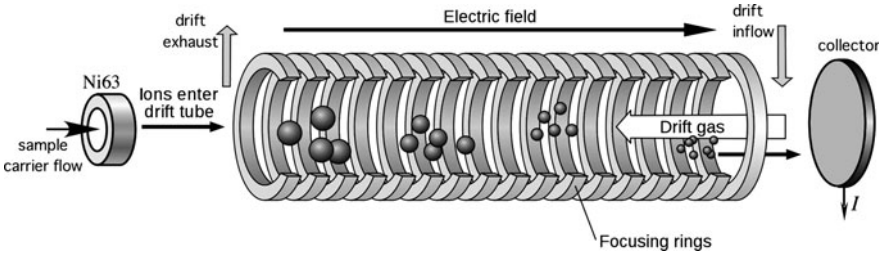

Fig. 17.18 Principle of ion mobility spectrometry

By increasing the deflection voltage of the electric field, all ion beams are successively directed onto the collector electrode, where the ion current I is measured. Differentiating the recorded $I(V)$ curve results in the ion mobility spectrum. Under constant conditions, the ion mobility K is a characteristic measure for a certain ion species. Typically, an IMS will have a high resolution of $R > 20$ (17.14):

$$R = t_d / W_{t,1/2} \tag{17.14}$$

where t_d is the drift time and $W_{t,1/2}$ the temporal peak width measured at half of the maximum peak height.

Ion mobility spectrometers have become common in security and screening applications, such as in airports where they are used for drug [65], and explosives detection [66]. Such systems have been made into hand-held sniffers and benchtop instruments.

Research and development of this detection technology remains very active with several groups around the world dedicated to advancing this analytical approach. Variations include different methods of ionization, the addition of chemicals to enhance ionization, and alternating electric fields to improve ion separation [67].

17.4.4.4 Thermal Sensors

When the internal energy of a system changes, it is accompanied by an absorption or evolution of heat (as defined by the first law of thermodynamics). Therefore, a chemical reaction, which is associated with heat, can be detected by an appropriate thermal sensor, such as those described in Chap. 16. These sensors operate on the basic principles that form the foundation of a microcalorimetry. The operating principle of a thermal sensor is simple: A temperature probe is coated with a chemically selective layer. Upon the introduction of a sample, the probe measures transfer of heat during the reaction between the sample and the coating.

A simplified drawing of such a sensor is shown in Fig. 17.19. It contains a thermal shield to reduce heat loss to the environment and thermistor coated by a catalytic layer. The layer may be an enzyme immobilized into a matrix. An example of such a sensor is the enzyme thermistor using an immobilized glucose oxidase (GOD). The enzymes are immobilized on the tip of the thermistor, which is then enclosed in a glass jacket in order to reduce heat loss to the surrounding solution. Another similar sensor with similarly immobilized bovine serum albumin is used as a reference. Both thermistors are connected as the arms of a Wheatstone bridge [68]. The temperature increase, dT, as a result of a chemical reaction is proportional to the incremental change in the enthalpy dH

$$dT = \frac{1}{C_p} dH, \tag{17.15}$$

where C_p is the heat capacity.

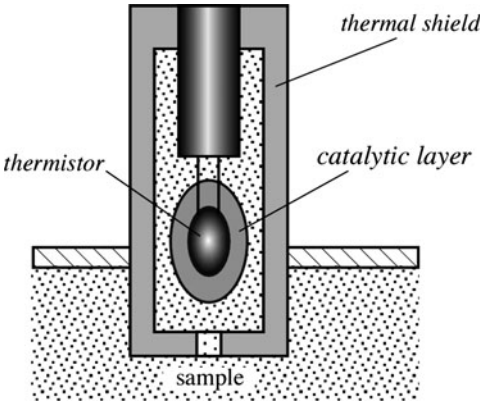

Fig. 17.19 Schematic diagram of a chemical thermal sensor

The chemical reaction in the coating is

$$\beta - \text{D-glucose} + H_2O + O_2 \xrightarrow{\text{GOD}} H_2O + \text{D-gluconic acid}, \Delta H_1, \tag{17.16}$$

and

$$H_2O_2 \xrightarrow{\text{catalase}} \frac{1}{2}O_2 + H_2O, \Delta H_2, \tag{17.17}$$

where ΔH_1 and ΔH_2 are partial enthalpies, the sum of which (for the above reaction) is approximately -80 kJ/mol. The sensor responds linearly with the dynamic range depending on the concentration of hydrogen peroxide (H_2O_2).

17.4.4.5 Pellister Catalytic Sensors

Pellisters and other catalytic sensors operate on the principle similar to thermal enzymatic sensors. Heat is liberated as a result of a catalytic reaction taking place at the surface of the sensor and the related temperature change inside the device is measured. On the other hand, the chemistry is similar to that of high temperature metal-oxide sensors. Catalytic gas sensors have been designed specifically to detect low concentrations of flammable gases in ambient air inside mines. These sensors often are called pellisters [69]. The platinum coil is imbedded in a pellet of ThO_2/Al_2O_3 coated with a porous catalytic metal: palladium or platinum (Fig. 17.20). The coil acts as both the heater and the resistive temperature detector (RTD). Naturally, any other type of heating element and temperature sensor can be successfully employed. When the combustible gas reacts at the catalytic surface, the heat evolved from the reaction increases the temperature of the pellet and of the platinum coil, thus increasing its resistance. There are two possible operating modes of the sensor. One is isothermal, where an electronic circuit controls the current through the coil to maintain a constant temperature. In the nonisothermal

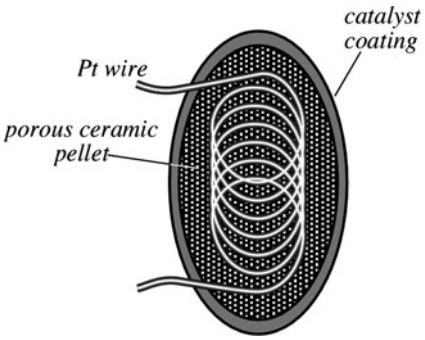

Fig. 17.20 Pellister or catalytic type detector

mode, the sensor is connected as a part of a Wheatstone bridge whose output voltage is a measure of the gas concentration.

17.4.5 Optical Transducers

Optical transducers measure the interactions of various forms of light or electromagnetic radiation and a target chemical or a selective layer, by detecting the modulation of some properties of the radiation. Examples of such modulations are variations in intensity, polarization, and velocity of light in a medium. The presence of different chemicals in the analyte affects which wavelengths of light are modulated. Optical modulation is studied by spectroscopy, which provides information on various microscopic structures from atoms to the dynamics in polymers. In a general arrangement, the monochromatic radiation passes through a sample (which may be gas, liquid, or solid), and its properties are examined at the output. Alternatively, the sample may respond with a secondary radiation (e.g. induced luminescence), which is also measured.

17.4.5.1 Infrared Detection

Most chemicals can absorb infrared light at wavelengths representative of the types of bonds present. For these chemicals, the Lambert–Beer law can be used because the absorbance of the gas is proportional to the concentration. Most small portable systems employing this technology use Nondispersive IR (NDIR). In NDIR, a polychromatic light source, typically a lamp or LED, is used to pass electromagnetic energy through a gas sample (Fig. 17.21a). Gases may be sampled using a fan or pump, or simply allowed to diffuse through a filter into an optically transparent cell, as is typically done in CO_2 sensors of this type. An optical filter is used in front of the light detector to limit the incoming light to only particular wavelengths

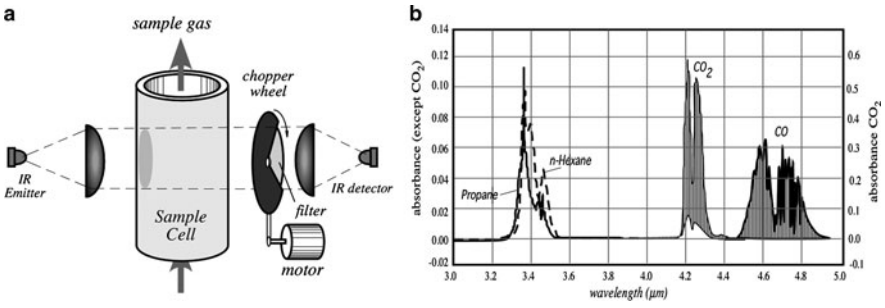

Fig. 17.21 Nondispersive IR (adapted from [70]) (**a**) and examples of absorption spectra (**b**)

associated with a target analyte (Fig. 17.21b). The attenuated absorption of that wavelength by the chemical indicates its presence and concentration.

Spectroscopic systems for measuring optical absorption are useful for UV and IR wavelengths and can be used to detect many chemicals by producing a more complex absorbance signature in the form of a spectrum. Benchtop IR instruments typically use dispersive IR techniques. In these instruments, a grate or prism is used to provide a broad wavelength range to select a specific wavelength of light to pass through the sample. In all strategies, the wavelength of the light source is routinely matched to the reactive energy of the optrode indicator to achieve a best possible electronic signal.

17.4.5.2 Fiber Optic Transducers

Fiber optical chemical sensors (Fig. 7.22) use a chemical reagent or sorbent phase to alter the amount or wavelength of light reflected by, absorbed by, or transmitted through a fiber waveguide (*see* also Fig. 4.17A). A fiber optic sensor typically contains three parts, a source of incident (pilot) light, an optrode, and a transducer (detector) to convert the changing photonic signal to an electrical signal. It is the optrode that contains the reagent phase membrane or indicator whose optical properties are affected by the analyte [71].

The location of the reagent and the specific optical characteristic that is affected by it vary from one type of optical sensor to another. Simple polymer-coated fibers coat the polished lens end of a glass fiber with a reagent that absorbs incident light. Coating the cladding of a fiber instead of its polished end affects the reflection and refraction of the light. This is referred to as evanescent wave sensing. While the glass optical fiber is rugged and in many cases chemically resistant, the coating or indicator is not and becomes the weak component in the system [72].

Differential designs (to isolate all but reaction of interest) are often employed to split the original incoming light source and pass one through the reagent area while the other is unaltered. The two optical paths are either multiplexed to a single detector (transducer) or fed to different transducers to produce a difference signal

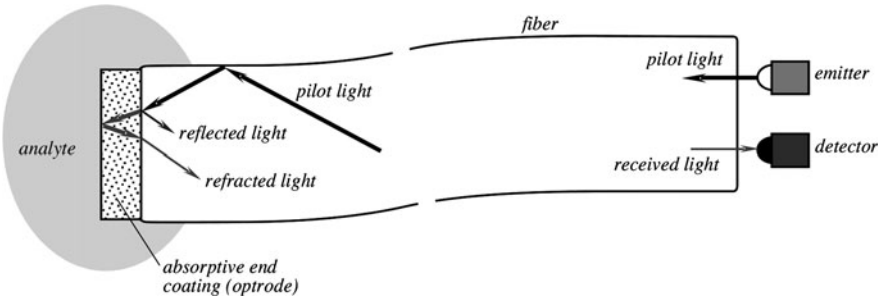

Fig. 7.22 Fiber optic gas sensor

used for sensing. One variation of fiber optic sensors is the use of coated beads, which are attached or embedded into the end or a surface of an optical fiber [73]. These beads can be modified or coated to have chemical or biological sensitivity.

17.5 Biochemical Sensors

Biosensors are a special class of chemical sensors. Evolution of species by means of natural selection led to extremely sensitive organs, which can respond to presence of just few molecules. Man-made sensors, while generally not as sensitive, employ biologically active materials in combination with several physical sensing elements, for example, amperometric or thermal, as described earlier. The biorecognition element is actually a bioreactor on the top of the conventional sensor, so the response of the biosensor will be determined by the diffusion of the analyte, reaction products, coreactants or interfering species, and the kinetics of the recognition process. Examples of biological elements that may be detected qualitatively and quantitatively by the biosensors are organisms, tissues, cells, organelles, membranes, enzymes, receptors, antibodies, and nucleic acids [74].

In fabrication of a biosensor, one of the key issues is immobilization of biochemical transducers on the physical or electrical transducer. The immobilizing layer or surface must confine the biologically active material on a sensing element and keep it from leaking out over the lifetime of the biosensor, allow contact to the analyte solution, allow any product to diffuse out of the immobilization layer, and not denature the biologically active material. Most of the biologically active materials used in biosensors are proteins or contain proteins in their chemical structures. To immobilize the proteins on the surface of the sensor, two basic techniques are employed: binding or physical retention. Adsorption and covalent binding are the two types of binding techniques. The retention involves separating the biologically active material from analyte solution with a layer on the surface of the sensor, which is permeable to the analyte and any products of the recognition reaction, but not to the biologically active material.

17.5.1 Enzyme Sensors

One of the most efficient ways of achieving selectivity is by using sensors with enzymatic layers. Enzymes are a special kind of catalyst, proteins of molecular weight 6–4,000 kDa found in living organisms. They have two remarkable properties: (1) they are extremely selective to a given substance and (2) they are extraordinarily effective in increasing the rate of reactions. Therefore, they favorably contribute to both the selectivity and the magnitude of the output signal. The maximum velocity of the reaction is proportional to the concentration of the enzyme. A general diagram of an enzymatic sensor is shown in Fig. 17.23 [74].

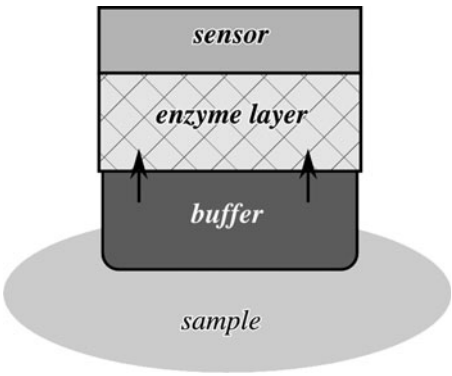

Fig. 17.23 Schematic diagram of an enzyme sensor

The sensing element can be a heated probe,[6] an electrochemical sensor, or an optical sensor. Enzymes operate only in an aqueous environment, so they are incorporated into immobilization matrices that are gels, specifically hydrogels. One common mode of operation is as follows. An enzyme (a catalyst) is immobilized inside a layer into which the target chemical from the sample diffuses. When the enzyme reacts with the target chemical, it produces a chemical product or other effect, which can be detected. Any other species that participates in the reaction must also diffuse in and out of these layers.

17.6 Multisensor Arrays

Processing multiple measurements from individual chemical sensors, and from a number of different or independent sensors, can provide information needed to statistically reduce error and improve both selectivity and sensitivity of a chemical sensor [75] or chemical detection instrument. Since measurement error is a sum of systematic error and random error, the measurement error of an individual sensor can be statistically reduced via multiple samples by using statistics to reduce or eliminate the random error [76]. Multiple redundant sampling can provide enough data to reduce the measurement standard deviation by a factor of $1/\sqrt{n}$, where n is the number of redundant samples. The redundant samples may come from the same sensor, or multiple sensors of the same type, to further insure the best possible response [77]. This, however, is useful against random errors but is not efficient against systematic errors.

Responses from multiple independent sensors of different types can be combined (often referred to as sensor fusion) to provide overlapping reinforced responses that better span the sensors' response spaces leaving fewer gaps where analyte identification would be weak or unavailable.

[6] See the subsection Thermal Sensors above.

17.7 Electronic Noses and Tongues

As two of the basic senses of humans, smell and taste play a very important role in daily life for recognizing environmental conditions. Electronic smell and taste sensors (e-noses and e-tongues) have been intensely researched and developed due to their broad potential commercial applications. They are very useful in the food industry, environmental protection, medicine, military, and other areas. The detection ability of these systems mainly depends on the ability of the sensitive materials to absorb or react with specific odors and ions. Although some achievements have been made, the e-noses and e-tongues still have significant limitations in sensitivity and specificity, compared with the biology binding of specific odorants and tastants to the olfactory and taste receptor cells.

A natural olfactory system is specialized to detect small airborne molecules at concentrations as low as a few parts per trillion and to discriminate among myriads of distinct compounds. In mammals, odorant molecules in the air enter the nostrils and bind with sensory neurons in the nose that convert the chemical interactions into an electrical signal that the brain interprets as a smell. The extraordinary performances of the olfactory and gustatory sensing are due to numerous receptors that are the sensory neurons and their subsequent neuronal processing. Each individual neuron can bind to multiple distinct molecules and ions with distinct affinities and specificities, though some receptors are relatively restricted to a set of few chemically related compounds. In humans, there are about 350 types of sensory neurons and many variations. Animals, such as dogs have hundreds more types of sensory neurons than humans. The process of olfactory perception begins when volatile compounds approach (via the respiratory air-stream) the nasal neuroepithelium where millions of distinct olfactory sensory cells reside.

Figure 17.24 shows a simplified diagram of the olfactory neuron. The perception occurs at tiny hair-like protrusions (cilias) from the receptor cells. The cilias are covered by mucus that captures the odorant molecules. The molecular properties of odorants that provide sensory properties are low-water solubility, high-vapor pressure, low polarity, ability to dissolve in fat (lipophilicity), and surface activity. Electrical response of the receptor cell is transmitted via axon to the next level of signal processing. Nearly, all chemical sensors suffer from a relatively short life due to contamination of the sensing

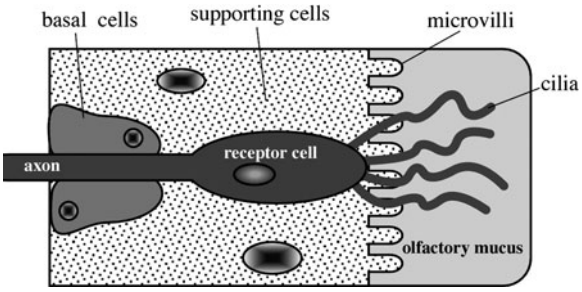

Fig. 17.24 Olfactory neuron

element. Nature solved this problem by frequent regeneration of all olfactory neurons that are replaced about every 40 days in humans, a very rare case of neural regeneration.

In contrast with traditional chemical sensors, receptors in the nasal cavities of mammals do not detect individual chemicals selectively, but use thousands of partially selective receptors that absorb inhaled chemicals. Each partially selective receptor may respond to a specific odorant strongly, weakly, or not at all, resulting in a distinct pattern that is sent to the brain for interpretation. The brain determines if this "smell" pattern has been detected before (learning and memory) and associates the chemical with a specific odor. Thus, different analytes give the brain different patterns, and these patterns determine the perception of the odor. It should be stressed that recognition of odorant species is not a job for any specific "olfactory sensor." The odor or taste sensation and recognition are accomplished by a much more complex system (brain) where the sensor is an integral part. A nose and tongue are not just sensors but rather extensions of the brain.

Odor and taste recognition is a sequential process where the molecule identification is gradual by narrowing down a selection by means of a layered pattern recognition technique – from a coarse to fine signature of the odorant molecule. Generally, the recognition is faster and the sensitivity is better for more complex compounds. For example, there is a remarkable regularity in the sensitivity values for hydrocarbon chains (alkanols) such as methanol, ethanol, propanol, butanol, and pentanol. As the length of the chain increases, the sensitivity increases [78]. For a chain of eight carbon atoms (octanol), the human sensitivity is about 10 ppb (part per billion), while for one carbon atom (methanol) it is 1 ppt (part per thousand).

The sensation of taste is also initiated by the interaction of tastants with receptors and ion channels in the apical microvilli of taste receptor cells when some sapid molecules dissolve in saliva [79]. Subsequently, through a cellular signaling pathway, gustatory signals are transduced gradually into the brain that integrates and analyzes these signals. Utilizing olfactory and gustatory cells as sensitive materials to develop a bioelectronic nose or tongue chip is one of the independent trends concerning the research and development of e-noses and e-tongues.

Much like people detect and remember many different smells and tastes and then use that knowledge to generalize about the perceptions they have not encountered before, the e-nose and e-tongue also need to be trained to recognize the chemical signatures of different smells before it can deal with unknowns. The current trend is a bionic approach based on combination of the multiple sensors and a signal processing by neural networks or their computer-based equivalents. The main idea is to use many sensors of different types and process data in a way that resembles data processing by living brains [80]. Electronic noses and tongues are less of a sensor or instrument and more of a measurement strategy.

The sensing parts of the e-noses and e-tongues are devices consisting of many similar, yet different sensing elements. This multisensory approach was successfully implemented at NIST.[7] The technology is based on interactions between chemical

[7]U.S. National Institute of Standards and Technology. www.nist.gov

species and semiconducting sensing materials placed on top of MEMS microheater platforms developed at NIST [81, 82, 83]. The NIST electronic nose was comprised of eight types of sensors in the form of metal-oxide films deposited on the surfaces of 16 microheaters, with two copies of each material (Fig. 17.25). A polycrystalline silicon resistor was used for heating. A thermal time constant of the heater is a couple of millisecond. Chemically sensitive films were deposited on the top of the heater: tin (SnO_2), tin oxide coated with titanium oxide (SnO_2/TiO_2), titanium oxide (TiO_2), and titanium oxide coated with ruthenium oxide ($TiO_2/RuOx$). These oxides are known to undergo chemical interactions with gas species ranging from surface-mediated oxidation of analyte gases to charge transfer upon analyte chemisorption. Precise control of the individual heating elements allowed treating each of them as a collection of "virtual" sensors at 350 temperature increments between 150 and 500°C, increasing the number of sensors to about 5,600. The combination of the

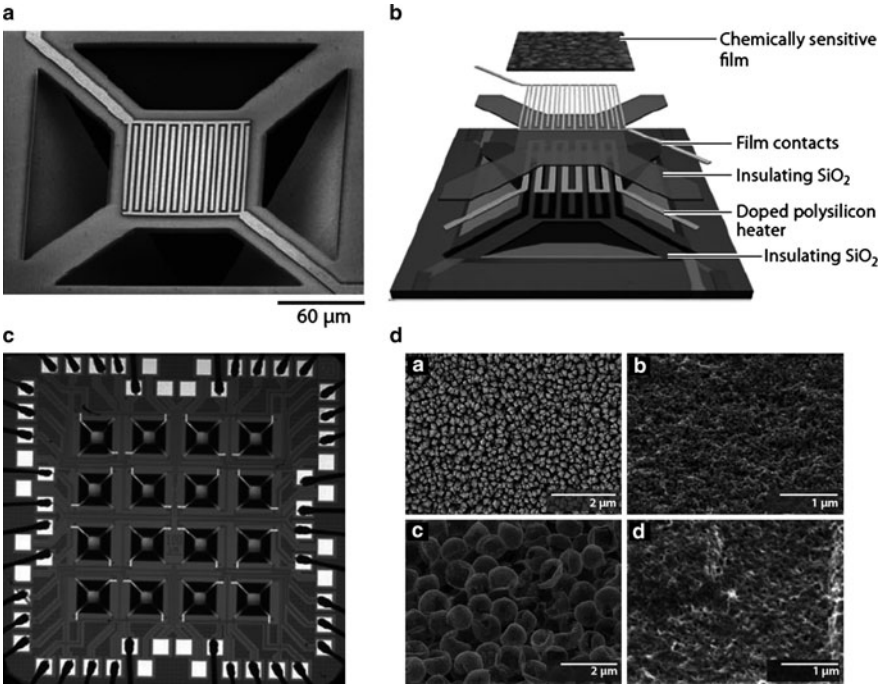

Fig. 17.25 Top view of a single chemical-sensor element (**a**); expanded schematic of the critical components of the microsensor device (**b**); a 16-element microsensor array (**c**); and scanning electron microscope images of a variety of chemically sensitive films deposited on microhot-plate-sensor platforms; and (**d**) images are shown at different scales to highlight each film's nanostructure and morphology: (*a*) Polycrystalline metal-oxide films deposited via chemical vapor deposition. Shown here is tin oxide (SnO_2); titanium oxide (TiO_2), TiO_2 layered on SnO_2, and ruthenium (Ru) deposited on TiO_2 have also been used. (*b*) Mesoporous TiO_2 film applied by drop coating. (*c*) Drop-coated shells of Sb:SnO_2. (*d*) Electrophoretically deposited nanostructured conductive polymer (here, colloidal olyaniline). Courtesy of Dr. Steve Semancik, NIST (stephen.semancik@nist.gov)

sensing films and the ability to vary the temperature gave the device the analytical equivalent of a snoot full of sensory neurons.

17.8 Specific Difficulties

The difficulty of developing chemical sensors (and systems) vs. other sensors (such as temperature, pressure, humidity, etc.) is that interactions with chemicals during the sensing process can result in permanent changes in the sensor. This typically results in drift in the sensor baseline which can adversely affect sensor calibration. For example, electrochemical cells, which employ liquid electrolytes (material that conducts electrical current via charged ions, not electrons) consume a small amount of electrolyte with each measurement requiring that the electrolyte be replenished eventually, chemical FET sensors may build up carbonic acid at the gate-membrane interface, which etches its components, and absorbent polymer coatings can become oxidized in harsh environments.

Also, unlike pressure or temperature sensors that have comparatively few conditions under which they need to be modeled to operate, chemical sensors are often exposed to nearly unlimited numbers of chemical combinations. This introduces interference responses, for example many chemical sensors have some degree of sensitivity to water. Therefore, when developing a sensor system to operate in the environment, the operator must account for changes in humidity when calibrating the system.

Ceramic bead-type and other catalytic hydrocarbon sensors begin to sinter ($\sim$400°C), and bulk platinum electrodes and heating elements begin to evaporate at elevated (1,000°C) temperatures, limiting their life spans and their usefulness for long-term continuous monitoring [84]. This evaporation rate is even higher in the presence of combustible gases. The loss of the platinum metal results in a change in the resistance of the wire that introduces offset error into the sensor reading, and leads to early burnout of the heating platinum coil.

Chemical poisoning can affect many sensors such as the catalytic bead devices where, chlorinated, sulfur and lead containing compounds can irreversibly bind to the sensing element, inhibiting the oxidation of the hydrocarbon species, and producing an inaccurate false-low reading. Filters are commonly used with any chemical sensor if it is to be subjected to an environment containing a characteristic poison. Judicious selection of the filter material is required to eliminate only the poisoning agent without an associated reduction in the target analyte (the chemical species being exposed to the sensor).

Surface acoustical wave devices that use species-selective adsorptive films can be poisoned mechanically by species that adsorb, but which do not desorb returning the mass of the device back to its original (calibrated) state. Similarly, gas-selective coatings on fiber-optic devices also may be poisoned by nonremovable species, permanently reducing the optical reflectance and indicating a false-positive.

Another problem unique to chemical sensors is the significant chemical reaction changes that occur at different the concentration levels. For example, certain reactive hydrocarbon devices (metal-oxide devices, voltammetric devices, etc.) require mixtures near stoichiometric (balanced chemical reactions) so that required minimal levels of both target analyte hydrocarbons and needed oxygen are available to feed the measurement reaction. If the hydrocarbon levels are too high (or better stated as the accompanying oxygen levels are too low) then only a fraction of the hydrocarbons will react producing a false negative reading.

References

1. Jacoby M (2009) Keepers of the gate. *Chem Engng News* 87(22):10–13
2. Zheng O, Noll RJ, Cooks RG (2009) Handheld miniature ion trap mass spectrometers. *Anal Chem* 81(7):2421–2425
3. Nagle HT, Gutierrez-Osuna R, Schiffman SS (1998) The how and why of electronic noses. *IEEE Spectrum* 35:22–34
4. Amoore JE, Johnston JW, Rubin M (1964) The stereochemical theory of odor. *Sci Am* 210:42–99
5. Ho CK, Hughes RC (2002) In-situ chemiresistor sensor package for real-time detection of volatile organic compounds in soil and groundwater. *Sensors* 2:23–34
6. Kim T (2009) Canary in the old growth. *High Country News*, Paonia, Colorado, February 16
7. For a wealth of information on Mine Safety Gas Monitoring Equipment is the United States Department of Labor. Mine Safety & Health Administration (MSHA) website: http://www.msha.gov
8. Clutton-Brock J (1995) In: Serpell J (ed) *The domestic dog, its evolution, behaviour and interactions with people*, Cambridge University Press, Cambridge, pp 7–20
9. Madou MJ, Morrison SR (1989) *Chemical sensing with solid state devices*, Academic Press, New York
10. Wolfrum EJ, Meglen RM, Peterson D, Sluiter J (2006) Metal oxide sensor arrays for the detection, differentiation, and quantification of volatile organic compounds at sub-parts-per-million concentration levels. *Sens Actuators* B 115:322–329
11. Persaud K, Dodd G (1982) Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* 299:352–355
12. Sberveglieri G, Kluwer (ed) (1992) *Gas sensors: principles, operations, and developments*, Academic Publishers, Boston, MA, pp 8, 148, 282, 346–408
13. Blum LJ (1997) *Bio- and chemi-luminescent sensors*, World Scientific, River Edge, NJ, pp 6–32
14. Sberveglieri G (1995) Recent developments in semiconducting thin-film gas sensors. *Sens Actuators* B 23:103–109
15. Demarne V, Sanjinés R (1992) Thin film semiconducting metal oxide gas sensors. In: G. Sberveglieri (ed) *Gas sensors*, Kluwer Academic, Dordrecht, Netherlands, pp 89–116
16. Malyshev VV, Vasiliev AA, Eryshkin AV, Koltypin EA, Shubin YI, Buturlin AI, Zaikin VA, Chakhunashvili GB (1992) Gas sensitivity of SnO_2 and ZnO thin-film resistive sensors to hydrocarbons, carbon monoxide, and hydrogen. *Sens Actuators* B 10:11–14
17. Hoefer U, Kühner G, Schweizer W, Sulz G, Steiner K (1994) CO and CO_2 thin-film SnO_2 gas sensors on Si substrates. *Sens Actuators* B 22:115–119
18. Demarne V, Grisel A (1998) An integrated low-power thin-film CO gas sensors on silicon. *Sensors and Actuators* B 13:301–313

19. Barsan N, Tomescu A (1995) The temperature dependence of the response of SnO_2-based gas sensing layers to O_2, CH_4, and CO. *Sens Actuators* B 26–27:45–48
20. Van Geloven P, Moons J, Honore M, Roggen J (1989) Tin (IV) oxide gas sensors: thick-film versus metallo-organic based sensors. *Sens Actuators* B 17:361–368
21. Schierbaum KD, Geiger J, Weimar U, Göpel W (1993) Specific palladium and platinum doping for SnO_2-based thin film sensor arrays. *Sens Actuators* B 13–14:143–147
22. Sulz G, Kuhner G, Reiter H, Uptmoor G, Schweizer W, Low H, Lacher M, Steiner K (1993) Ni, In, and Sb implanted Pt and V catalyzed thin-film SnO_2 gas sensors. *Sens Actuators* B 16:390–395
23. Tournier G, Pijolat C, Lalauze R, Patissier B (1995) Selective detection of CO and CH_4 with gas sensors using SnO_2 doped with palladium. *Sens Actuators* B 26–27:24–28
24. Huck R, Böttger U, Kolh D, Heiland G (1993) Spillover effects in the detection of H_2 and CH_4 by sputtered SnO_2 films with Pd and PdO deposits. *Sens Actuators* B 17:355–359
25. Saji K, Takahashi H, Kondo H, Takeuchi, Igarashi I (1983) Characteristics of TiO_2 oxygen sensor in nonequilibrium gas mixtures. In: Seiyama T, Fueki K, Shiokawa J, Suzuki S (eds) Chemical sensors, proceedings of the international meeting on chemical sensors, Fukuoka Japan, Elsevier, Tokyo, pp 171–176
26. Mumuera G, Gonzalez-Ellpe AR, Munoz A, Fernandez A, Soria J, Conesa J, Sanz J (1989) Mechanism of hydrogen gas-sensing at low temperatures using Rh/TiO_2 Systems. *Sens Actuators* B 18:337–348
27. Egashira M, Kanehara N, Shimizu Y, Iwanaga H (1989) Gas-sensing characteristics of Li+-doped and undoped ZnO whiskers. Sens Actuators B 18:349–360
28. Gentry SJ (1988) Catalytic devices. In: Edmonds TE (ed) *Chemical sensors*. Chapman and Hall, New York
29. Cobbold RSC (1974) *Transducers for biomedical measurements*. Wiley, New York
30. www.askiitians.com/iit-jee-chemistry/physical-chemistry/Kohlrausch-law.aspx
31. Tan TC, Liu CC (1991) Principles and fabrication materials of electrochemical sensors. *Chemical sensor technology*. 3, Kodansha Ltd
32. Clark LC (1956) Monitor and control of blood and tissue oxygen tension. *Trans Am Soc Artif Internal Organs* 2:41–46
33. Grate JW, Klusty M, Barger WR, Snow AW (1990) Role of selective sorption in chemiresistor sensors for organophosphorus detection. *Anal Chem* 62(18):1927–1934
34. Ho CK, Hughes RC (2002) In-situ chemiresistor sensor package for real-time detection of volatile organic compounds in soil and groundwater, *Sensors* 2:23–34
35. Hierlemann A, Lange D, Hagleitner C, Kerness N, Koll A, Brand O, Baltes H (2000) Application-specific sensor systems based on CMOS chemical microsensors. *Sens Actuators B Chem* 70:2–11
36. Endres H-E, Hartinger R, Schwaiger M, Gmelch G, Roth M (1999) A capacitive CO_2 sensor system with suppression of the humidity interference. *Sens Actuators B Chem* 57:83–87
37. Patel SV, Mlsna TE, Fruhberger B, Klaassen E, Cemalovic S, Baselt DR (2003) Chemicapacitive microsensors for volatile organic compound detection. Sens Actuators B 96(3):541–553
38. Fotis E (2002) A new ammonia detector based on thin film polymer technology. *Sensors* 19 (5):73–75
39. Mlsna TE, Cemalovic S, Warburton M, Hobson ST, Mlsna DA, Patel SV (2006) Chemicapacitive microsensors for chemical warfare agent and toxic industrial chemical detection. *Sens Actuators B Chem* 116(1–2):192–201
40. The Multi-User MEMS Process (MUMPs) from MEMSCAP, Inc. (Durham, NC) is used to manufacture the these chemicapacitive sensor chips.
41. Britton CL, Jones RL, Oden PI, Hu Z, Warmack RJ, Smith SF, Bryan WL, Rochelle JM (2000) Multiple-input microcantilever sensors. *Ultramicroscopy* 82:17–21
42. Baselt DR, Fruhberger B, Klaassen E, Cemalovic S, Britton CL, Patel SV, Mlsna TE, McCorkle D, Warmack Jr, B (2003) Design and performance of a microcantilever-based hydrogen sensor, *Sens Actuators B Chem* 88(2):120–131

43. Polk BJ (2002) ChemFET arrays for chemical sensing microsystems, *IEEE* 732–735
44. Wróblewski W, Wojciechowski K, Dybko A, Brzózka Z, Egberink RJM, Snellink-Ruël BHM, Reinhoudt DN (2001) Durability of phosphate-selective CHEMFETs, *Sens Actuators B: Chem* 78(1–3):315–319
45. Wilson DM, Hoyt S, Janata J, Booksh K, Obando L (2001) Chemical sensors for portable, handheld field instruments, *IEEE Sensor J* 1(4):256–274
46. Janata J (1989) *Principles of chemical sensors*, Chapter 4. Plenum Press, New York
47. Kharitonov AB, Zayats M, Lichtenstien A, Katz E, Willner I (2000) Enzyme monolayer-functionalized field-effect transistors for biosensor applications. *Sens Actuators B* 70 (1–3):222–231
48. Ballantine DS, White RM, Martin SJ, Ricco AJ, Frye GC, Zellers ET, Wohltjen H (1997) *Acoustic wave sensors: theory, design and physicochemical applications*, Academic Press, Boston, MA
49. Ristic VM (1983) Principles of acoustic devices. Wiley, New York
50. Nieuwenhuizen MS et al (1986) Transduction mechanism in SAW gas sensors. *Electron Lett* 22:184–185
51. Wenzel SW, While RM (1989) Analytic comparison of the sensitivities of bulk-surface-, and flexural plate-mode ultrasonic gravimetric sensors. *Appl Phys Lett* 54:1976–1978
52. Nieuwenhuizen MS et al (1986) Transduction mechanism in SAW gas sensors. *Electron Lett* 22:184–185
53. Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. *Phys Rev Lett* 56:930–933
54. Battiston FM, Ramseyer J-P, Lang HP, Baller MK, Gerber Ch, Gimzewski JK, Meyer E, Guntherodt H-J (2001) A chemical sensor based on a microfabricated cantilever array with simultaneous resonance-frequency and bending readout, *Sens Actuators B Chem* 77:122–131
55. Baselt DR, Fruhberger B, Klaassen E, Cemalovic S, Britton Jr, CL, Patel SV, Mlsna TE, McCorkle D, Warmack B (2003) Design and performance of a microcantilever-based hydrogen sensor. *Sens Actuators B* 88(2):120–131
56. Hansen KM, Ji, H-F, Wu G, Datar R, Cote R, Majumdar A, Thundat T (2001) Cantilever-based optical deflection assay for discrimination of DNA single-nucleotide mismatches. *Anal Chem* 73:1567–1571
57. Baselt DR, Lee GU, Natesan M, Metzger SW, Sheehan PE, Colton RJ (2001) A biosensor based on magnetoresistance technology. *Biosens Bioelectron* 13:731–739
58. Betts TA, Tipple CA, Sepaniak MJ, Datskos PG (2000) Selectivity of chemical sensors based on micro-cantilevers coated with thin polymer films. *Anal Chim Acta* 422:89–99
59. Senesac LR, Yi D, Greve A, Hales JH, Davis ZJ, Nicholson DM, Boisen A, Thundat T. (2009) Micro-differential thermal analysis detection of adsorbed explosive molecules using micro-fabricated bridges. *Rev Sci Instrum* 80:035102
60. Thundat T, Wachter EA, Sharp SL, Warmack RJ (1995) Detection of mercury-vapor using resonating microcantilevers. *Appl Phys Lett* 66(13):1695–1697
61. Thundat T, Chen GY, Warmack RJ, Allison DP, Wachter EA (1995) Vapor detection using resonating microcantilevers. *Anal Chem* 67(3):519–521
62. Pinnaduwage LA, Wig A, Hedden DL, Gehl A, Yi D, Thundat T, Lareau RT (2004) Detection of trinitrotoluene via deflagration on a microcantilever, *J Appl Phys* 95:5871–5875
63. Datskos PG, Oden PI, Thundat T, Wachter EA, Warmack RJ, Hunter SR (1996) Remote infrared radiation detection using piezoresistive microcantilevers, *Appl Phys Lett* 69 (20):2986–2988
64. Creaser C, Thomas P et al. (2004) Ion mobility spectrometry: a review. Part 1. Structural analysis by mobility measurement. *The Analyst* 129:984–994
65. Ching W, William FS, Herbert HH Jr (2000) Secondary electrospray ionization ion mobility spectrometry/mass spectrometry of illicit drugs. *Anal Chem* 72(2):396–403
66. Maggie T, Herbert HH Jr (2004) Secondary electrospray ionization-ion mobility spectrometry for explosive vapor detection. *Anal Chem* 76(10):2741–2747
67. Rhykerd CL, Hannum DW, Murray DW, Parmeter JE (1999) Guide for the Selection of Commercial Explosives Detection Systems for Law Enforcement Applications, NIJ Guide

100–99, NCJ 178913, September 1999, available at: www.ojp.usdoj.gov/nij/pubs-sum/178913.htm
68. Dewa AS, Ko WH (1994) Biosensors. In: Sze SM (ed) *Semiconductor sensors*, Wiley, New York, pp 415–472
69. Gentry SJ (1988) Catalytic devices. In: Edmonds TE (ed) *Chemical sensors*, Chapman and Hall, New York
70. RAE Systems Inc., Theory and Operation of NDIR Sensors, Technical Note TN-169. rev 1 wh.04-02
71. Dybko A, Wroblewski W (2000) Fiber optic chemical sensors, www.ch.pw.edu.pl/~dybko/csrg/fiber/operating.html
72. Seiler K, Simon W (1992) Principles and mechanisms of ion-selective optodes. *Sensors Actuators B* 6:295–298
73. Walt DR (2000) Molecular biology: bead based fiber-optic arrays. *Science* 287(5452):451
74. Dewa AS, Ko WH (1994) Biosensors. In: Sze SM (ed) *Semiconductor sensors*, Wiley, Inc. New York, pp 415–472
75. Gottuk DT, Hill SA, Schemel CF, Strehlen BD, Rose-Pehrsson SL, Shaffer RE, Tatem PA, Williams FW (1999) Identification of Fire Signatures for Shipboard Multi-criteria Fire Detection Systems. NRL/MR/6180-99-8386, Naval Research Laboratory, Washington, DC, pp 48–87
76. Einax JW, Zwanziger HW, Geib S (1997) *Chemometrics in environmental analysis*. VCH, Weinheim, Germany, pp 2–75
77. Prasad L, Iyengar SS, Rao RL, Kashyap RL (1994) Fault-tolerant sensor integration using multiresolution decomposition. *Phys Rev E* 49(4):3452–3461
78. Cometto-Muñiz JE, Cain WS (1990) Thresholds for odor and nasal pungency. *Physiol Behav* 48:719–725
79. Wang P, Liu Q, Xua Y, Cai H, Li Y (2007) Olfactory and taste cell sensor and its applications in biomedicine. *Sens Actuators A* 139:131–138
80. Nagle HT, Schiffman SS, Gutierrez-Osuna R (1998) The how and why of electronic noses, *IEEE Spectrum* 35:22–34
81. Raman B, Meier DC, Evju JK, Semancik S (2009) Designing and optimizing microsensor arrays for recognizing chemical hazards in complex environments. *Sens Actuators* B 137:617–629
82. Raman B, Hertz JL, Benkstein KD, Semancik S (2008) Bioinspired methodology for artificial olfaction. *Anal Chem* 80:8364
83. Meier DC, Raman B, Semancik S (2009) Detecting chemical hazards with temperature-programmed microsensors: overcoming complex analytical problems with multidimensional databases. *Annu Rev Anal Chem* 2:463–84
84. Edmonds TE (ed) (1988) *Chemical sensors*, Blackie and Son Ltd, New York

Chapter 18
Sensor Materials and Technologies

Any sufficiently advanced technologyis in distinguishable from magic.

– Arthur C. Clarke

Methods of sensor fabrication are numerous and specific for each particular design. They comprise processing of semiconductors, optical components, metals, ceramics, and plastics. Here, we briefly describe some materials and the most often used techniques.

18.1 Materials

18.1.1 Silicon as Sensing Material

Silicon is present in the sun and stars and is a principal component of a class of meteorites known as aerolites. Silicon is the second most abundant material on Earth, being exceeded only by oxygen – it makes up to 25.7% of the earth's crust, by weight. Silicon is not found free in nature but occurs chiefly as the oxide and as silicates. Some oxides are sand, quartz, rock crystal, amethyst, clay, mica, etc. Silicon is prepared by heating silica and carbon in an electric furnace using carbon electrodes. There are also several other methods for preparing the element. Crystalline silicon has a metallic luster and grayish color.[1] The Czochralski process is commonly used to produce single crystals of silicon used for the solid-state semiconductors and micromachined sensors. Silicon is a relatively inert element, but it is attacked by halogens and diluted alkali. Most acids, except hydrofluoric, do not affect it. Elemental silicon transmits infrared radiation and is commonly used as windows and lenses in the mid- and far-infrared sensors.

[1]Silicon should not be confused with *silicone* that is made by hydrolyzing *silicon* organic chloride such as dimethyl silicon chloride. Silicones are used as insulators, lubricants, and for production of silicone rubber.

J. Fraden, *Handbook of Modern Sensors*,
DOI 10.1007/978-1-4419-6466-3_18, © Springer Science+Business Media, LLC 2010

Silicon's atomic weight is 28.0855, and its atomic number is 14. Its melting point is 1,410°C and boiling point is 2,355°C. Specific gravity at 25°C is 2.33 and valence is 4.

Properties of silicon are well studied and its applications to sensor designs have been extensively researched around the world. The material is inexpensive and can now be produced and processed controllably to unparalleled standards of purity and perfection. Silicon exhibits a number of physical effects that are quite useful for sensor applications (Table 18.1).

Unfortunately, silicon does not posses the piezoelectric effect (or perhaps fortunately, because in many sensors piezoelectricity would generate interferences). Most effects of silicon, such as the Hall effect, the Seebeck effect, the piezoresistance, etc. are quite large; however, a major problem with silicon is that its responses to many stimuli show substantial temperature sensitivity. For instance, strain, light, and magnetic field responses are temperature-dependent. When silicon does not display the proper effect, it is possible to deposit layers of materials with the desired sensitivity on top of the silicon substrate. For instance, sputtering of ZnO thin films is used to form piezoelectric transducers that are useful for the fabrication of surface acoustic waves (SAW) devices and accelerometers. In the later case, the strain at the support end of an etched micromechanical cantilever is detected by a ZnO overlay.

Silicon itself exhibits very useful mechanical properties, which nowadays are widely used to fabricate such devices as pressure transducers, temperature sensors, and force and tactile detectors by employing the microelectromechanical systems (MEMS) technologies . Thin film and photolithographic fabrication procedures make it possible to realize a great variety of extremely small, high-precision mechanical structures using the same processes that have been developed for electronic circuits. High-volume batch-fabrication techniques can be utilized in the manufacture of complex, miniaturized mechanical components that may not be possible with other methods. Table A.14 in Appendix presents a comparative list of mechanical characteristics of silicon and other popular crystalline materials.

Although single-crystal silicon (SCS) is a brittle material, yielding catastrophically (not unlike most oxide-based glasses) rather than deforming plastically (like most metals), it certainly is not as fragile as is often believed. The Young's modulus of silicon (1.9×10^{12} dyne/cm or 27×10^{6} psi) has a value of that approaching

Table 18.1 Effects in the silicon-based sensors [1]

Stimuli	Effects
Radiant	Photovoltaic effect , photoelectric effect, photoconductivity , and photomagneto-electric effect
Mechanical	Piezoresistivity , lateral photoelectric effect, and lateral photovoltaic effect
Thermal	Seebeck effect , temperature dependence of conductivity and junction, and Nernst effect
Magnetic	Hall effect , magneto-resistance, and Suhi effect
Chemical	Ion-sensitivity

stainless steel and is well above that of quartz and most of glasses. The misconception that silicon is extremely fragile is based on the fact that it is often obtained in thin slices (5–13 cm diameter wafers), which are only 250–500 μm thick. Even stainless steel at these dimensions is very easy to deform inelastically.

As mentioned earlier, many of the structural and mechanical disadvantages of SCS can be alleviated by the deposition of thin films. Sputtered quartz, for example, is utilized routinely by industry to passivate integrated circuit chips against airborne impurities and mild atmospheric corrosion effects. Another example is a deposition of silicon nitrate (Table A.14), which has hardness second only to diamond. Anisotropic etching is the key technology for micromachining of miniature three-dimensional structures in silicon. Two etching systems are of practical interest. One based on ethylenediamine and water with some additives. The other consists of purely inorganic alkaline solutions like KOH, NaOH, or LiOH.

Forming the so-called polysilicon (PS) materials allows to develop sensors with unique characteristics. Polysilicon layers (on the order of 0.5 μm) may be formed by vacuum deposition onto oxided silicon wafer with an oxide thickness of about 0.1 μm [2]. Polysilicon structures are doped with boron by a technique known in the semiconductor industry as low-pressure chemical vapor deposition (LPCVD).

Figure 18.1a shows resistivity of boron-doped LPCVD polysilicon in comparison with SCS. Resistivity of PS layers is always higher than that of a single crystal material, even when the boron concentration is very high. At low doping concentrations, the resistivity climbs rapidly, so that only the impurity concentration range

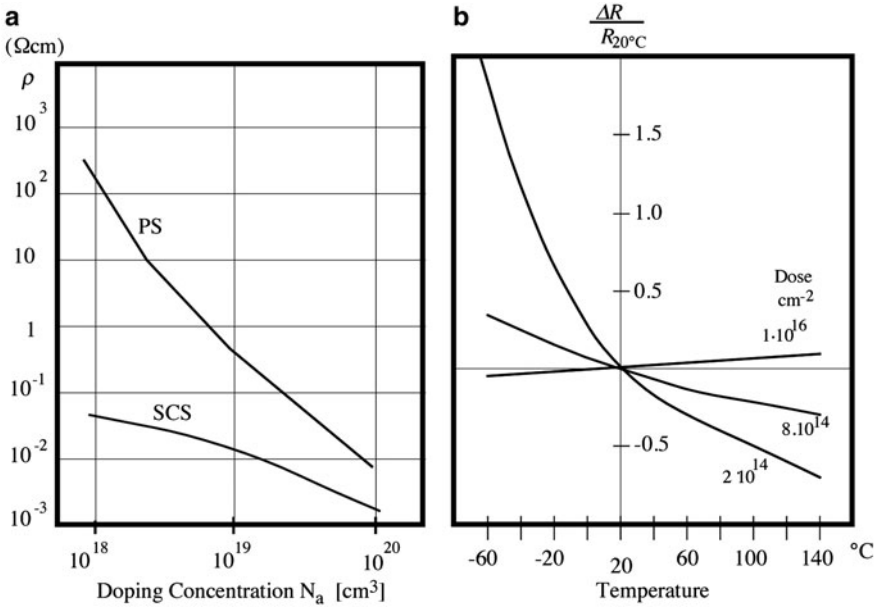

Fig. 18.1 Specific resistivity of boron-doped silicon (**a**) and temperature coefficient of resistivity of silicon for different doping concentrations (**b**)

is of interest to a sensor fabrication. The resistance change of PS with temperature is not linear. The temperature coefficient of resistance may be selected over a wide range, both positive and negative, through selected doping (Fig. 18.1b). Generally, the temperature coefficient of resistance increases with decreased doping concentration. Resistance at any given temperature of a PS layer may be found from

$$R(T) = R_{20} e^{\alpha_R (T - T_0)}, \tag{18.1}$$

where $\alpha_R = \frac{1}{R_{20}} \frac{dR(T_0)}{dT}$ is the temperature coefficient, and R_{20} is the resistance at calibrating point (T_0=20°C). Figure 18.2a shows that the temperature sensitivity of PS is substantially higher than that of SCS and can be controlled by doping. It is interesting to note that at a specific doping concentration, the resistance becomes insensitive to temperature variations (point Z).

For the development of sensors for pressure, force, or acceleration, it is critical to know the strain sensitivity of PS resistors expressed through the gauge factor. Figure 18.2b shows curves of the relative resistance change of boron dopes PS resistors, referenced to the resistance value R_0 under no-stress conditions, as a function of longitudinal strain ε_1. The parameter varies with the implantation dose. It can be seen that the resistance decreases with compression and increases under tension. It should be noted that the gauge factor (the slope of the line in Fig. 18.2b) is temperature-dependent. PS resistors are capable of realizing at least as high a

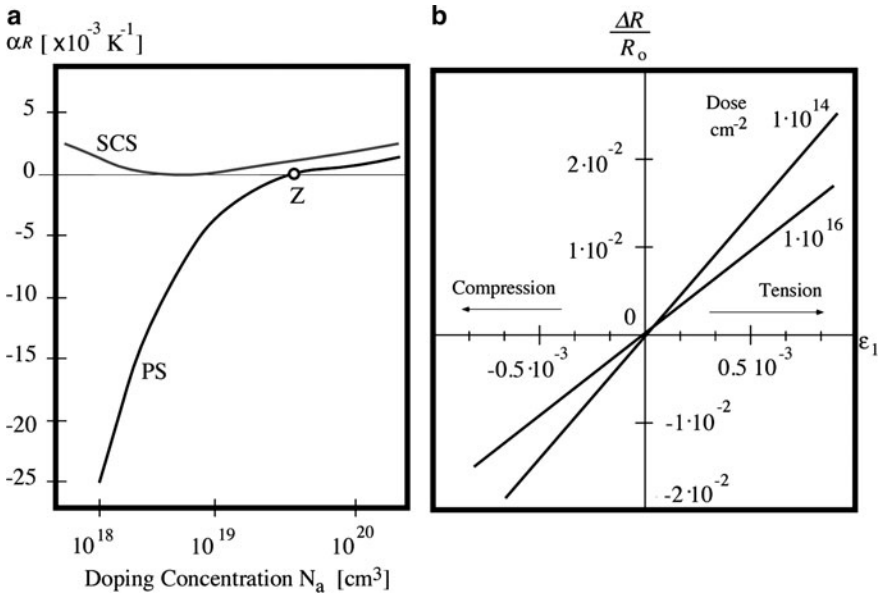

Fig. 18.2 Temperature coefficient as function of doping (**a**) and piezoresistive sensitivity of silicon (**b**)

level of long-term stability as any can be expected from resistors in SCS, since surface effects play only a secondary role in device characteristics.

18.1.2 Plastics

Plastics are synthetic materials made from chemical raw materials called monomers. A monomer (one chemical unit) such as ethylene is reacted with other monomer molecules to form long chains of repeating ethylene units, forming the polymer polyethylene. In a similar manner, polystyrene is formed from styrene monomers. The polymers consist of carbon atoms in combination with other elements. Polymer chemists use only eight elements to create thousands of different plastics. These elements are carbon (C), hydrogen (H), nitrogen (N), oxygen (O) fluorine (F), silicon (Si), sulfur (S), and chlorine (Cl). Combining these elements in various ways produce extremely large and complex molecules.

Each atom has a limited capacity (energy bonds) for joining with other atoms, and every atom within a molecule must have all its energy bonds satisfied if the compound is to be stable. For example, hydrogen can bond only to one other atom, while carbon or silicon must attach to four other atoms to satisfy its energy bonds. Thus, H–H and H–F are stable molecules, while C–H and Si–Cl are not. Figure 18.3 shows all eight atoms and the corresponding energy bonds.

Adding more carbon atoms in a chain and more hydrogen atoms to each carbon atom creates heavier molecules. For example, ethane gas (C_2H_6) is heavier than methane gas because it contains additional carbon and two hydrogen atoms. Its molecular weight is 30. Then, the molecular weight can be increased in the increments of 14 (one carbon + two hydrogen), until the compound pentane (C_5H_{12}) is reached. It is too heavy to be gas and indeed it is liquid at room temperature. Further additions of CH_2 groups make progressively heavier liquid

Element	Atomic weight	Energy Bonds	
Hydrogen	1	–H	1
Carbon	12	–C–	4
Nitrogen	14	–N–	3
Oxygen	16	–O–	2
Fluorine	19	–F	1
Silicon	28	–Si–	4
Sulfur	32	–S–	2
Chlorine	35	–Cl	1

Fig. 18.3 The atomic building blocks for polymers

until $C_{18}H_{38}$ is reached. It is solid-paraffin wax. If we continue and grow larger molecules, the wax becomes harder and harder. At about $C_{100}H_{202}$, the material with a molecular weight 1,402 is tough enough and is called a low-molecular weight polyethylene the simplest of all thermoplastics. Continuing the addition of more CH_2 groups further increases the toughness of the material until medium molecular weight (between 1,000 and 5,000 carbons) and high-molecular-weight polyethylene. Polyethylene, being the simplest polymer (Fig. 18.4), has many useful properties in the sensor technologies. For example, the polyethylene is reasonably transparent in the mid- and far-infrared spectral ranges and thus is used for the fabrication of infrared windows and Fresnel lenses.

By applying heat, pressure, and catalysts, monomers are grown into long chains. The process is called polymerization. Chain length (molecular weight) is important because it determines many properties of a plastic. The major effects of increased length are increased toughness, creep resistance, stress-crack resistance, melt temperature, melt viscosity, and difficulty of processing. After polymerization is completed, the finished polymer chains resemble long intertwined bundles of spaghetti with no physical connections between chains. Such a polymer is called thermoplastic (heat-moldable) polymer.

If chains are packed closer to one another, more denser polyethylene is formed that in effect results in formation of crystals. The crystallized areas are stiffer and stronger. Such polymers are more difficult to process since they have higher and sharp melt temperatures. That is, instead of softening, they quickly transform into low-viscosity liquids. On the other hand, amorphous thermoplastics soften gradually, but they do not flow as easily as crystalline plastics. The examples of

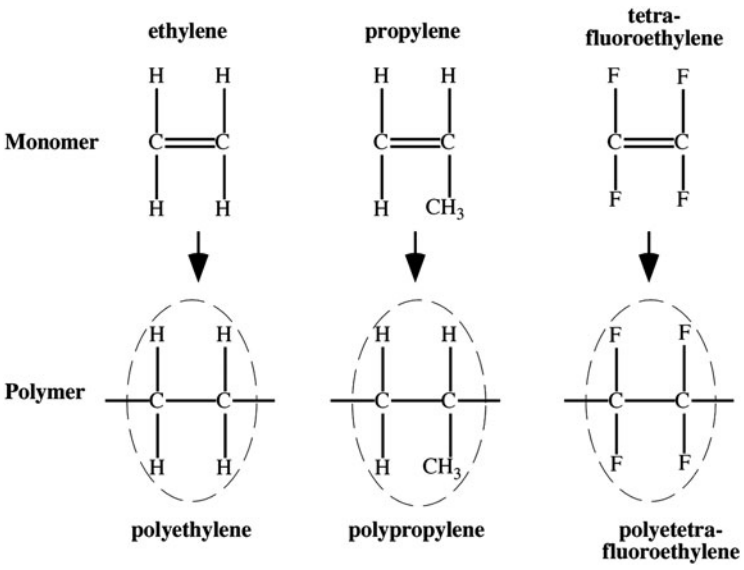

Fig. 18.4 Monomers and their respective polymer units

amorphous polymers are ABS, polystyrene, polycarbonate, polysulfone, and polyetherimide. Crystalline plastics include polyethylene, polypropylene, nylon, PVDF, acetal, and others.

Following is a list of nonexhaustive thermoplastics.

Acrylonitrile–butadiene–styrene (*ABS*) is very tough, yet hard and rigid. It has fair chemical resistance, low-water absorption, and good dimensional stability. Some grades may be electroplated.

Acrylic has high optical clarity and excellent resistance to outdoor weathering. This is hard, glossy material with good electrical properties. It is available in a variety of colors.

Fluoroplastics comprise a large family of materials (PTFE, FEP, PFA, CTFE, ECTFE, ETFE, and PFDF) with excellent electrical properties and chemical resistance, low friction, and outstanding stability at high temperatures. However, their strength is moderate and cost is high. One example is polytetrafluoroethylene (PTFE), which is known as DuPont brand Teflon.

Nylon (polyimide) has outstanding toughness and wear resistance with low coefficient of friction. It has good electrical and chemical properties. However, it is hygroscopic and dimensional stability is worst than in most other plastics.

Polycarbonate has the highest impact resistance. It is transparent with excellent outdoor stability and resistance to creep under load. It may have some problems with chemicals. Polyester has excellent dimensional stability but is not suitable for outdoor use or for service in hot water.

Polyethylene is lightweight and inexpensive with excellent chemical stability. It has good electrical properties. Moderate transparency in broad spectral range from visible to far infrared. It has poor dimensional and thermal stability.

Polypropylene has outstanding resistance to flex and stress cracking with excellent chemical and electrical properties with good thermal stability. It is lightweight and cheap. Optical transparency is good down to far-infrared spectral range. However, absorption and scattering of photons in mid-infrared range is higher than in polyethylene.

Polyurethane is tough and extremely abrasion- and impact-resistant. It can be made into films and foams. It has good chemical and electrical properties; however, UV exposure degrades its quality.

Polybutadiene is a synthetic rubber that has a high resistance to wear. It has been used to coat or encapsulate electronic assemblies, offering extremely high electrical resistivity. It exhibits a recovery of 80% after stress is applied.

Polyvinyl chloride (*PVC*) is the third most widely used thermoplastic polymer after polyethylene and polypropylene. PVC is cheap, durable, and easy to assemble. It can be made softer and more flexible by the addition of plasticizers, the most common being phthalates. In electronics, it is used to make flexible tubing and electrical cable insulation.

Other type of plastics is called *thermoset* in which polymerization (curing) is done in two stages: one by the material manufacturer and the other by the molder. An example is phenolic, which during the molding process is liquefied under pressure, producing a cross-linking reaction between molecular chains. After it

has been molded, a thermoset plastic has virtually all its molecules interconnected with strong physical bonds, which are not heat reversible. In effect, curing a thermoset is like cooking an egg. Once it is cooked, it will remain hard. In general, thermoset plastics resist higher temperatures and provide greater dimensional stability. This is the reason why such thermoset plastics as polyester (reinforced) used to make boat hulls and circuit-breaker components, epoxy is used to make printed circuit boards, and melamine is used to make dinnerware. On the other hand, thermoplastics offer higher impact strength, easier processing, and better adaptability to complex designs than do thermosets.

The thermoplastics that are most useful in sensor-related applications are the following.

Alkyd has excellent electrical properties and very low moisture absorption.

Allyl (diallyl phthalate) has outstanding dimensional stability and high heat and chemical resistance.

Epoxy has exceptional mechanical strength, electrical properties, and adhesion to most of materials.

Phenolic is a low-cost material. Color is limited to black and brown.

Polyester (thermoplastic version) has a great variety of colors, may be transparent or opaque. Shrinkage is high.

If two different monomers (A and B) are combined in a polymerization reaction, such a polymer is called *copolymer*. Final properties of copolymer depends on ratio of components A and B.

Polymer mechanical properties can be modified by providing additives such as fibers to increase strength and stiffness, plastisizers for flexibility, lubricants for easier molding, or UV stabilizers for better performance in sun light.

Another good way to control properties of plastics is to make polymer alloys or blends. Primarily, this is done to retain properties of each component.

Conductive plastics. Being wonderful electrical isolators, plastic materials often require lamination with metal foil, painting with conductive paint, or metallization to give them electrical conductive properties, required for shielding. Another way of providing electrical conductivity is mixing plastics with conductive additives (for instance, graphite or metal fibers) or building composite plastic parts incorporating metal mesh.

Piezoelectric plastics are made from PVF_2, PVDF, and copolymers that are crystalline materials. Initially, they do not possess piezoelectric properties and must be poled either in high voltage or by corona discharge (Sect. 3.6.2). Metal electrodes are deposited on both sides of the film either by silk-screening or by vacuum metallization. These films, in some applications are used instead of ceramics, thanks to their flexibility and stability against mechanical stress. Another advantage of the piezoelectric plastics is their ability to be formed into any desirable shape.

A polymer that is very useful for the sensing technologies is Kapton, which is a *polyimide* (PI) film developed by DuPont. It is a thermoset material with density of 1.42 g/cm^3 and low-thermal conductivity of 0.12 W/(mK). PI can remain stable in a wide range of temperatures from near the absolute zero of -273°C to $+400^\circ$C

(≈0–673 K). Among other things, PI is used in flexible printed circuits that can be employed to connect sensors to rigid printed circuit boards (PCB). A flexible PI printed circuit board can be made as thin as 50 μm and even thinner. These boards are flexible and can withstand close to a million bends. PI is also commonly used as a material for windows of all kinds at X-ray sources (synchrotron beam-lines and X-ray tubes) and X-ray detectors. Its high mechanical and thermal stability as well as its high transmittance to X-rays makes it the preferred material. It also does not suffer from radiation damage. However, PI has relatively poor resistance to mechanical abrasion.

18.1.3 Metals

From the sensor designer standpoint, there are two classes of metals: nonferrous and ferrous. Ferrous metals, like steel, are often used in combination with magnetic sensors to measure motion, distance, magnetic field strength, etc. Also, they are quite useful as magnetic shields. Nonferrous metals, on the other hand, are permeable to magnetic fields and used whenever these fields are of no concern.

Nonferrous metals offer a wide variety of mechanical and electrical properties. When selecting a metal, one must consider not only its physical properties, but also ease of mechanical processing. For example, copper has excellent thermal and electrical properties, yet it is difficult to machine, so in many instances aluminum should be considered as a compromise alternative.

Aluminum has a high strength-to-weight ratio and possesses its own anticorrosion mechanism. When exposed to air, aluminum does not oxide progressively, like iron would do. The protection is provided by a microscopic oxide coating that forms on the surface and seals the bare metal from environment.

There are hundreds of aluminum alloys. They can be processed by many ways, like drawing, casting, stamping. Some alloys can be soldered and welded. Besides excellent electrical properties, aluminum is a superb reflector of light over nearly entire spectrum from UV to radio waves. Aluminum coatings are widely used for mirrors and waveguides. In the mid- and far-infrared range, the only superior to aluminum reflector is gold.

Beryllium has several remarkable properties. Its low density (two thirds that of aluminum) is combined with high modulus per weight (five times that of steel), high specific heat, excellent dimensional stability, and transparency to X-rays. However, this is a expensive metal. Like aluminum, beryllium forms a protective coating on its surface, thus resisting to corrosion. It may be processed by many conventional methods, including powder cold pressing. The metal is used as X-ray windows, optical platforms, mirror substrates, and satellite structures.

Magnesium is a very light metal with high strength-to-weight ratio. Due to its low modulus of elasticity, it can absorb energy elastically, which gives its good damping characteristics. The material is very easy to process by most of metal working techniques.

Nickel allows designing of very tough structures that are also resistant to corrosion. When compared with steel, the nickel alloys have ultrahigh strength and high modulus of elasticity. Its alloys include binary systems with copper, silicon, and molybdenum. Nickel and its alloys preserve their mechanical properties down to cryogenic temperatures and at high temperatures up to 1,200°C. Nickel is used in high-performance superalloys such as Inconell, Monel (Ni–Cu), Ni–Cr, and Ni–Cr–Fe alloys.

Copper combines very good thermal and electrical conductivity properties (second only to pure silver) with corrosion resistance and relative ease of processing. However, its strength-to-weight ratio is relatively poor. Copper is also difficult to machine. Copper and its alloys – the brasses and bronzes come in variety of forms, including films. Brasses are alloys that contain zinc and other designated elements. Bronzes comprise main groups: copper-tin-phosphorus (phosphor bronze), copper-tin-lead-phosphorus (lead phosphor bronzes), and copper-silicon (silicon bronzes) alloys. Under the outdoor condition, copper develops a blue-green patina. This can be prevented by applying acrylic coating. Copper alloy with beryllium has excellent mechanical properties and used to make springs.

Lead is the most impervious of all common metals to X-rays and γ-radiation. It resists attack by many corrosive chemicals, most types of soil, marine, and industrial environment. It has low melting temperature, ease of casting and forming, and good sound and vibration absorption. It possesses natural lubricity and wear resistance. Lead is rarely used in pure form. Its most common alloys are "hard lead" (1–13% of antimony), calcium, and tin alloys that have better strength and hardness.

Platinum is a silver-white precious metal that is extremely malleable, ductile, and corrosion resistant. Its positive temperature coefficient of resistance is very stable and reproducible, which allows its use in temperature sensing.

Gold is extremely soft and chemically inert metal. It can only be attacked by aqua regia and by sodium and potassium in presence of oxygen. One gram of pure gold can be worked into a leaf covering 5,000 cm^2 and only less than 0.1 μm thick. Mainly, it is used for plating and alloyed with other metals like copper, nickel, and silver. In sensor applications, gold is used for fabricating electrical contacts and plating mirrors and wave-guides operating in the mid- and far-infrared spectral ranges.

Silver is least costly of all precious metals. It is very malleable and corrosion resistant. It has the highest electrical and thermal conductivity of all metals.

Palladium, *iridium*, and *rhodium* resemble and behave like platinum. They are used as electrical coatings to produce hybrid and printed circuit boards and various ceramic substrates with electrical conductors. Another application is in the fabrication of high quality reflectors operating in broad spectral range, especially at elevated temperatures or highly corrosive environments. Iridium has the best corrosion resistance of all metals and thus used in the most critical applications.

Molybdenum maintains its strength and rigidity up to 1,600°C. The metal and its alloys are readily machinable by conventional tools. In nonoxidizing environments, it resists attacks by most of acids. Its prime application is for high-temperature

devices such as heating elements and reflectors of intense infrared radiation for high-temperature furnaces. Molybdenum has low coefficient of thermal expansion and resists erosion by molten metals.

Tungsten in many respects is similar to molybdenum but can operate even at higher temperatures. A thermocouple sensor fabricated of tungsten is alloyed with 25% of rhenium with another wire-with 5% of rhenium.

Zinc is seldom used alone, except for coating, and mainly used as an additive in many alloys.

18.1.4 Ceramics

In sensor technologies, ceramics are very useful crystalline materials thanks to their structural strength, thermal stability, light weight, resistance to many chemicals, ability to bond with other materials, and excellent electrical properties. Although most metals form at least one chemical compound with oxygen, only a handful of oxides are useful as the principal constituent of ceramics. Examples are alumina and beryllia. The natural alloying element in the alumina is silica; however, alumina can be alloyed also with chromium, magnesium, calcium, and other elements.

Several metal carbides and nitrades qualify as ceramics. Most commonly used are boron carbide and nitrate and aluminum nitrade (Table A.24). Whenever fast heat transfer is of importance, aluminum nitrade shall be considered, while silicon carbide has high dielectric constant, which makes it attractive for designing capacitive sensors. Due to their hardness, most ceramics require special processing. A precise and cost-effective method of cutting various shapes of ceramic substrates is scribing, machining, and drilling by use of computer-controlled CO_2 laser. Ceramics for the sensor substrates are available from many manufacturers in thickness ranging from 0.1 to 10 mm.

18.1.5 Glasses

Glass is amorphous solid material made by fusing silica with a basic oxide. Although its atoms never arrange themselves into crystalline structure, atomic spacing in glass is quite tight. Glass is characterized by transparency, availability in many colors, hardness, and resistance to most chemicals except hydrofluoric acid (Table A.25). Most glasses are based on the silicate system and is made from three major components: silica (SiO), lime ($CaCO_3$), and sodium carbonite ($NaCO_3$). Nonsilicate glasses include phosphate glass (which resists hydrofluoric acid), heat absorbing glasses (made with FeO), and systems based on oxides of aluminum, vanadium, germanium, and other metals. An example of such specialty glass is arsenic trisulfate (As_2S_3) known as AMTIR, which is substantially transparent in

mid- and far-infrared spectral range and is used for fabricating infrared optical devices.

Borosilicate glass is the oldest type of glass that is substantially resistant to thermal shock. In the trademark Pyrex®, some of the SiO_2 molecules are replaced by boric oxide. This glass has a low coefficient of thermal expansion and thus is used for fabrication optical mirrors (like in telescopes).

Lead-alkali glass (lead glass) contains lead monoxide (PbO) that increases its index of refraction. Also, it is a better electrical insulator. In the sensor technologies, it is used for fabricating optical windows, prisms, and as a shield against nuclear radiation. Other glasses include alumosilicate glass (in which Al_2O_3 replaces some silica), 96% silica, and fused silica.

Another class of glass is *light-sensitive glasses* that are available in three grades. Photochromatic glass darkens when exposed to ultraviolet radiation and clears when the UV is removed or glass is heated. Some photochromatic compositions remain darkened for a week or longer. Others fade within few minutes, when UV is removed. The photosensitive glass reacts to UV in a different manner: If it is heated after exposure, it turns from clear to opal. This allows to create some patterns within the glass structure. Moreover, the exposed opalized glass is much more soluble in hydrofluoric acid, which allows for efficient etching technique.

18.1.6 Optical Glasses

18.1.6.1 Visible and Near Infrared Ranges

Most optical glasses are mixtures of silica obtained from beds of fine sand or from pulverized sandstone; an alkali to lower the melting point, usually a form of soda or, for finer glass, potash; lime as a stabilizer; and cullet (waste glass) to assist in melting the mixture. The properties of glass are varied by adding other substances, commonly in the form of oxides, e.g. lead, for brilliance and weight; boron, for thermal and electrical resistance; barium, to increase the refractive index; cerium, to absorb infrared rays; metallic oxides, to impart color; and manganese, for decolorizing.

Optical glasses are classified by their main chemical components and are identified by refractive index. Since refractive index is the function of a wavelength, it is measured at specific spectral lines of spectrum produced by various elements. Examples of the spectral lines and the corresponding refractive indices of some glasses and clear plastics are given in Table 18.2.

Quality and durability of glasses depend on the environment to which they are subjected. In various processes of fabricating optical components such as lenses and prisms, surface deterioration is often encountered and recognized as dimming, staining, and latent scratching. These surface defects are caused by chemical reactions of the glass constituents with water in the surrounding environment or

Table 18.2 Wavelengths of spectral lines and refractive indices for some glasses and plastics

Wavelength (nm)	Spectral line	Element	Glass BSC517642	Glass LAF744447	Plastic acrylic	Plastic polycarbonate
1,013.98	t	Hg	1.507	1.726		
852.11	S	Cs	1.510	1.730		
768.19	A'	K				
706.52	r	He				
656.27	C	H	1.514	1.739	1.489	1.578
643.85	C'	Cd				
632.8	632.8	He–Ne laser				
589.29	D	Na			1.492	1.584
587.56	d	He	1.516	1.744		
546.07	e	Hg	1.518	1.748		
486.13	F	H			1.498	1.598
479.99	F'	Cd	1.522	1.756		
435.83	g	Hg	1.526	1.765		
404.66	h	Hg	1.530	1.773		
365.01	i	Hg	1.536	1.787		

Note: Glasses from Pilkington Special Glass Ltd.

with detergents in the cleaning fluids. A high refractive index leads to weaker surfaces.

Polished glass exposed to high humidity and rapid temperature variations may "sweat". Water vapor may condense to form droplets on the glass surface. Some of the glass components that dissolve in the droplets may in turn attack the glass surface and react with gaseous elements in the air (e.g. CO_2). Reaction products form as white spots or a cloudy film as the glass surface dries. It is called "dimming". Water contact causes chemical reactions (ion exchange between cations in the glass and hydronium ions H_3O+ in water), which result in a silica-rich surface layer that causes an interference color on that layer. It is called "staining". Fine scratches created on the glass surfaces during polishing will sometimes grow to a large visible size when the surfaces are exposed to corrosive ions out of inorganic builders in a detergent used for cleaning.

18.1.6.2 Mid- and Far-Infrared Ranges

For operation in the range of thermal radiation (mid and far infrared), the silicon-based amorphous glasses have very high coefficient of absorption and thus cannot be used. The alternatives are the crystalline materials (e.g. germanium and silicon), some polymers (polyethylene and polypropylene), and special chalcogenide glasses based on use of selenium. These glasses can be drawn into fibers to form fiber optic sensors and thermal radiation transmission lines. To produce lenses and prisms, they can be molded, just like the silicon-based glasses or plastics. This can dramatically simplify production and lower cost when compared with the

Table 18.3 Properties of chalcogenide glasses (Courtesy of Amorphous Materials, Inc. Garland, TX)

	AMTIR-1	AMTIR-2	AMTIR-3	AMTIR-4	AMTIR-5	AMTIR-6	C1
Composition	Ge–As–Se	As–Se	Ge–Sb–Se	As–Se	As–Se	As–S	As–Se–Te
Transmission range (μm)	0.7–12	1.0–14	1.0–12	1.0–12	1.0–12	0.6–8	1.2–14
Refractive index at 10μm	2.4981	2.7613	2.6027	2.6431	2.7398	2.3807	2.8051
Upper use temperature (°C)	300	150	250	90	130	150	120

Table 18.4 Crystalline infrared materials

Material	Useful spectral range (μm)	Approximate refractive index
Magnesium fluoride	0.5–9.0	1.36
Zinc sulfide	0.4–14.5	2.25
Calcium fluoride	<0.4–11.5	1.42
Zinc selenide	0.5–22.0	2.44
Magnesium oxide	<0.4–9.5	1.69
Calcium telluride	0.9–31.0	2.70
Silicon	1.2–8.0	3.45
Germanium	1.3–22.0	4.00

crystalline materials that as a rule require grinding and polishing. Examples of the most popular chalcogenide glasses are given in Table 18.3.

An alternative to the AMTIR glasses is the use of crystalline materials having substantial transmission in the mid- and far-infrared ranges. The most popular IR materials and their properties are given in Table 18.4.

Note that a high refractive index in a glass results in a high coefficient of reflection. This may cause an undesirable loss in the signal intensity. To reduce reflection, antireflection (AR) coating is recommended.

In spite of their softness, plastics are popular materials for the mid- and far-infrared optics, thanks to very low cost, ease of molding, and low refractive index. The most common optical plastics are given in Table 18.2 where the refractive indices are close to those in the IR ranges.

18.1.7 Nanomaterials

Just few years ago, nanotechnology was a somewhat emotional term, more of a wishful thinking than a real thing. While it referred to dimensions of a device on a nanometer (10^{-9} m) scale, most of the subminiature elements had sizes about a thousand times larger – in a micrometer (10^{-6} m) range. Nowadays, this technology

progresses very rapidly and at the time of this writing, nanomaterials were produced on the scale of 10 nm.

One of the nanomaterials that is of a great interest for sensing (in a not-so-distant future) is the carbon nanotubes [7]. In a nanotube, molecules of carbon are arranged in a tubular shape with a remarkable length-to-diameter ratio of 28 million, which is significantly larger than any other known material. The diameter of a nanotube is just few nanometers, while the length is several millimeters. The tubes are characterized by extremely large surface area for a given volume. The tensile strength is over 100 times larger than that of stainless steel. They exhibit unique electrical properties and are efficient thermal conductors. The theoretical thermal conductivity of a nanotube is 20 times larger than that of copper – one of the best heat conductors. This makes the nanotube materials attractive for fabricating temperature and infrared sensors. At least in theory, the tubes can carry electric currents thousand times stronger than copper. Their final usage, however, may be limited by their potential toxicity.

18.2 Surface Processing

18.2.1 Deposition of Thin and Thick Films

Thin films are required to give a sensing surface some properties that it otherwise does not possess. For example, to enhance the absorption of thermal radiation by a far-infrared sensor, the surface may be coated with a material having high IR photon absorptivity, for instance nichrome. A piezoelectric film may be applied to a silicon waver to give it piezoelectric properties. The thick films are often used to fabricate pressure sensors or microphones where the flexible membranes have to be produced. Several methods may be used to deposit thin and relatively thin (often referred to as "thick") layers of films on a substrate or semiconductor wafer. Among them are the spin-casting, vacuum deposition, sputtering, electroplating, and screen printing [6].

18.2.2 Spin Casting

The spin-casting process involves use of a thin-film material dissolved in a volatile liquid solvent. The solution is pored on the sample and the sample is rotated in a high speed. The centrifugal forces spread the material and after the solvent evaporates, a thin layer of film remains on the sample. This technique is often used for deposition of organic materials, especially for fabricating humidity and chemical sensors. The thickness depends on solubility of the deposited material and the spin film and typically is in the range from 0.1 to 50 μm. Since the process relies on the

flow of the solution, it may not yield a uniform film or can form island (film-free areas) when the sample has a nonflat surface. Besides, the material may have tendency to shrinkage. Nevertheless, in many cases, it is a useful and often the only acceptable method of deposition.

18.2.3 Vacuum Deposition

A metal can be converted into gaseous form and then deposited on the surface of the sample. The evaporation system consists of a vacuum chamber (Fig. 18.5) where diffuse pump evacuates air down to 10^{-6} to 10^{-7} Torr of pressure. A deposited material is placed into a ceramic crucible that is heated by tungsten filament above the metal melting point. An alternative method of heating is the use of an electron beam.

On the command from the control device, the shutter opens and allows the metal atoms emanated from the molten metal to deposit on the sample. Parts of the sample that shall remain free of the deposited material are protected by the mask. The deposited film thickness is determined by the evaporation time and the vapor pressure of the metal. Hence, materials with low melting point are easy to deposit, for instance aluminum. In general, vacuum deposited films have large residual stress, and thus this technique is used mainly for depositing only thin layers.

Since the molten material is virtually a point source of atoms, it may cause both nonuniform distribution of the deposited film and the so-called shadowing effect

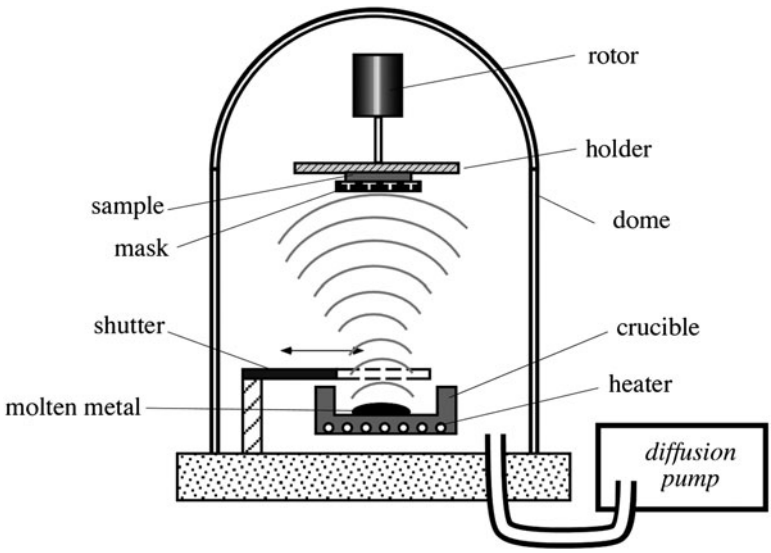

Fig. 18.5 Deposition of thin metal film in a vacuum chamber

where the edges of the masked pattern appear blurry. Two methods may help to alleviate this problem. One is use of multiple sources where more than one crucible (often 3 or 4) is used. Another method is rotation of the target.

When using the vacuum deposition, one shall pay attention to the introduction of spurious materials into the chamber. For instance, even minuscule amount of oil leaking from the diffuse pump will result in burning of organic materials and codeposition on the sample of such undesirable compounds as carbohydrates.

18.2.4 Sputtering

As in the vacuum deposition method, sputtering is performed in a vacuum chamber (Fig. 18.6); however, after evacuation of air, an inert gas, such as argon or helium is introduced into the chamber at about 2×10^{-6} to 5×10^{-6} Torr. An external high voltage dc or ac power supply is attached to the cathode (target), which is fabricated of the material that has to be deposited on the sample. The sample is attached to the anode at some distance from the cathode. High-voltage ignite plasma of the inert gas and the gas ions bombard the target. The kinetic energy of the bombarding ions is sufficiently high to free some atoms from the target surface. Hence, the escaped sputtered atoms deposit on the surface of the sample.

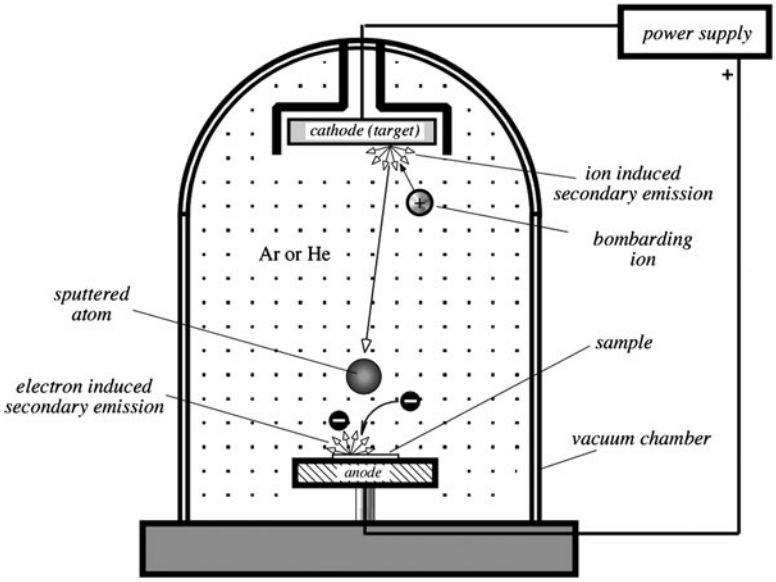

Fig. 18.6 Sputtering process in a vacuum chamber

The sputtered technique yields better uniformity, especially if magnetic field is introduced into the chamber allowing for better directing atoms toward the anode. Since this method does not require high temperature of the target, virtually any material, including organic, can be sputtered. Moreover, materials from more than one target can be deposited at the same time (co-sputtering), permitting controlled ratio of materials. For example, this can be useful for sputtering nichrome (Ni and Cr) electrodes on the surface of the pyroelectric sensors.

18.2.5 Chemical Vapor Deposition

A chemical vapor phase deposition (CVD) process is an important technique for the production of optical, optoelectronic, and electronic devices. For the sensor technologies, it is useful for forming optical windows and fabrication of semiconductor sensors where thin and thick crystalline layers have to be deposited on the surface. The CVD process takes place in a deposition (reaction) chamber, one of the versions of which in a simplified form is shown in Fig. 18.7.

The substrates or wafers are positioned on a stationary or rotating table (the substrate holder) whose temperature is elevated up to the required level by the heating elements. The top cover of the chamber has an inlet for the carrier H_2 gas, which can be added by various precursors and dopants. These additives, while being carried over the heated surface of the substrate, form a film layer. The gas mixture flows from the distribution cone over the top surface of the wafers and exits through the exhaust gas outlets. The average gas pressure in the chamber may by near 1 atm, or somewhat lower. For example, a layer 6,000 Å of Ga0.47 In0.53. As can be grown on the InP substrate at 1 atm and 630°C with a rate of 1.4 Å/s [3].

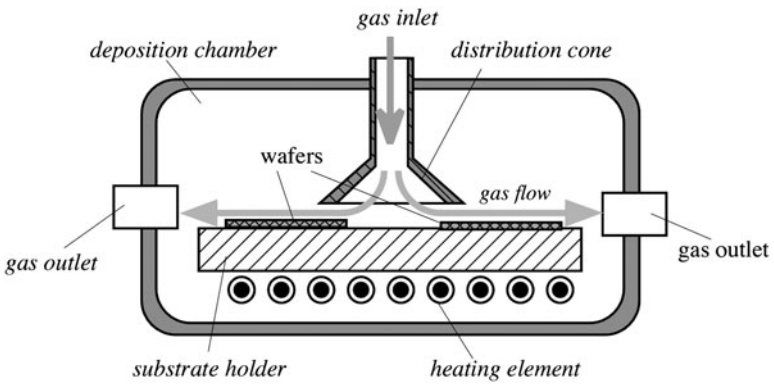

Fig. 18.7 Simplified structure of a CVD reactor chamber

18.2.6 Electroplating

Electroplating is the coating of an electrically conductive object with a layer of metal using electrical current. The result is a thin, smooth, even coat of metal on the object. Modern electrochemistry was invented by Italian chemist Luigi V. Brugnatelli in 1805. Brugnatelli used Alessandro Volta's invention of 5 years earlier, the voltaic pile, to facilitate the first electroplating.

The process used in electroplating is called electrodeposition and is analogous to a galvanic cell acting in reverse. The part to be coated is placed into a bath or tank containing a solution of one or more metal salts. The part that requires plating is connected to an electrical circuit, forming the cathode (negative) of the circuit while an electrode (typically of the same metal to be plated) forms the anode – a positive electrode. When an electrical current is passed through the circuit, metal ions in the solution take up excess electrons at the part.

The anode and cathode in the electroplating cell are connected to a dc power supply (Fig. 18.8). The metal of the anode is oxidized to form cations with a positive charge. These cations associate with the anions in the solution. The cations are reduced at the cathode to deposit in the metallic state.

A popular practical process that is different from the bath plating is called brush electroplating. The selected areas or entire part is plated using a brush saturated with the plating solution. The brush, typically a stainless-steel body wrapped with a cloth material that both holds the plating solution and prevents direct contact with the item being plated, is connected to the positive side of a low-voltage dc power supply. The part to be plated connected to the negative terminal. The brush acts as the anode, but typically does not contribute any plating material, although sometimes the brush is made from or contains the plating material in order to extend the life of the plating solution.

The plating most commonly is done by a single metallic element, not an alloy. However, some alloys can be electrodeposited, notably brass and tin/lead alloy.

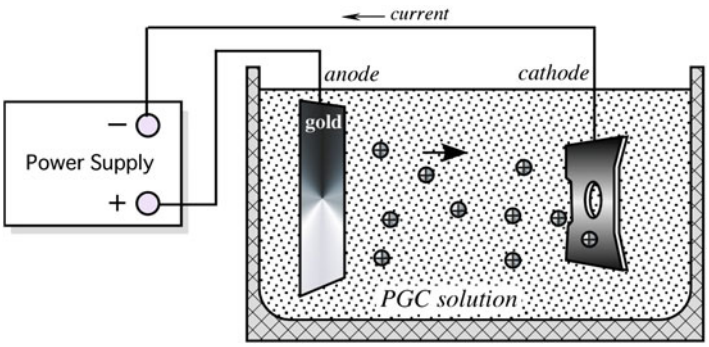

Fig. 18.8 Concept of bath electroplating (PGC means potassium gold cyanide salt used for gold plating)

Often, direct deposition of a metal on a part (substrate) is not the most efficient way of plating, mainly for the reliability reasons. Let us, for example, consider electroplating with a metal that has inherently poor adhesion to the substrate. In such a case, a "strike" (the under-plating) can be first deposited, The strike serves as a foundation for the subsequent plating processes. The strike is "friendly" or compatible with both the metal and the substrate. One example of this situation is a notably poor adhesion of electrolytic nickel on zinc alloys. The solution is to use the copper strike first, since copper has good adherence to most materials. A typical strike is a very thin (less than 0.1 μm thick) plating of an aid metal having high quality and good adherence to the plated material.

In the sensing technologies, one of the most frequently used metals for plating is gold. It serves to provide corrosion-resistant, electrically conductive layers on copper conductors and printed circuit boards, and also it is an excellent reflector for use in the mid- and far-infrared spectral ranges. However, plating gold directly on copper, if not done correctly, may pose serious problems because the copper atoms tend to diffuse through the gold layer, causing tarnishing of its surface and formation of an oxide and/or sulfide layer. In an IR reflector, this will result in the degradation of performance due to a dramatic reduction in reflectivity. A layer of a suitable barrier metal, usually nickel, is deposited on the copper substrate before the gold plating. The layer of underplating nickel provides mechanical backing for the gold layer, improving its wear resistance. It also reduces the impact of micropores that may be present in the gold layer.

18.3 Microtechnology

The present trend in the sensor technologies is undoubtedly shifted toward the microminiaturization or microsystem technologies, known as MST. A subset of these is known as MEMS [8]. A MEMS device has electrical and mechanical components, which means there must be at least one moving or deformable part and that electricity must be part of its operation. Another subset is called MEOMS that stands for microelectro-optical systems. As the name implies, at least one optical component is part of the device. Most of the sensors that are fabricated with use of MEMS or MEOMS are three-dimensional devices with dimensions in the order of micrometers.

The two constructional technologies of microengineering are microelectronics and micromachining. Microelectronics, producing electronic circuitry on silicon chips, is a very well-developed technology. Micromachining is the name for the techniques used to produce the structures and moving parts of microengineered devices. One of the main goals of microengineering is to be able to integrate microelectronic circuitry into micromachined structures to produce completely integrated systems (microsystems). Such systems typically have the same advantages of low cost, reliability, and small size as silicon chips produced in the microelectronics industry.

Presently, there are three micromachining techniques that are in use or are extensively developed by the industry [4, 5]. Silicon micromachining is given most prominence, since this is one of the better developed micromachining techniques. Silicon is the primary substrate material used in the production microelectronic circuitry, and so is the most suitable candidate for the eventual production of microsystems.

The excimer laser is an ultraviolet laser, which can be used to micromachine a number of materials without heating them, unlike many other lasers that remove material by burning or vaporizing it. The excimer laser lends itself particularly to the machining of organic materials (such as polymers).

The acronym LIGA comes from the German name for the process (Lithographie, Galvanoformung, Abformung). LIGA uses lithography, electroplating, and molding processes to produce microstructures.

18.3.1 Photolithography

Photolithography is the basic technique used to define the shape of micromachined structures in the three techniques outlined in the following paragraphs. The technique is essentially the same as that used in the microelectronics industry.

Figure 18.9A shows a thin film of some material (e.g. silicon dioxide) on a substrate of some other material (e.g. a silicon wafer). The goal of the process is to selectively remove some silicon dioxide (oxide) so that it only remains in particular areas on the silicon wafer (Fig. 18.9F). The process beings with producing a mask. This will typically be a chromium pattern on a glass plate. The wafer is then coated

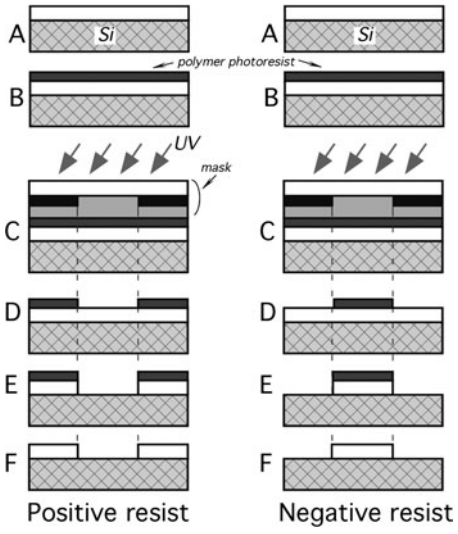

Fig. 18.9 Positive and negative photolithography

with a polymer that is sensitive to ultraviolet light (Fig. 18.9B), called a photoresist. Ultraviolet light is then shone through the mask onto the photoresist (Fig. 18.9C). The photoresist is then developed that transfers the pattern on the mask to the photoresist layer (Fig. 18.9D).

There are two types of photoresist, termed positive (lefts side of Fig. 18.9) and negative (right side of Fig. 18.9). Where the ultraviolet light strikes the positive resist, it weakens the polymer, so that when the image is developed, the resist is washed away where the light struck it – transferring a positive image of the mask to the resist layer. It is similar to glass-plate photography. The opposite occurs with the negative resist. Where the ultraviolet light strikes negative resist, it strengthens the polymer, so when developed, the resist that was not exposed to ultraviolet light is washed away – a negative image of the mask is transferred to the resist.

A chemical (or some other method) is then used to remove the oxide where it is exposed through the openings in the resist (Fig. 18.9E). Finally, the resist is removed leaving the patterned oxide (Fig. 18.9F).

18.3.2 Silicon Micromachining

There are a number of basic techniques that can be used to pattern thin films that have been deposited on a silicon wafer, to shape the wafer itself, and to form a set of basic microstructures (bulk silicon micromachining). The techniques for depositing and patterning thin films can be used to produce quite complex microstructures on the surface of silicon wafer (surface silicon micromachining). Electrochemical etching techniques are being investigated to extend the set of basic silicon micromachining techniques. Silicon bonding techniques can also be utilized to extend the structures produced by silicon micromachining techniques into multilayer structures.

18.3.2.1 Basic Techniques

There are three basic techniques associated with silicon micromachining. These are the deposition of thin films of materials, the removal of material (patterning) by wet chemical etchants, and the removal of material by dry etching techniques. Another technique that is utilized is the introduction of impurities into the silicon to change its properties (i.e. doping).

18.3.2.2 Thin Films

There are a number of different techniques that facilitate the deposition or the formation of very thin films (of the order of micrometers, or less) of different materials on a silicon wafer (or other suitable substrate). These films can then be

patterned using photolithographic techniques and suitable etching techniques. Common materials include silicon dioxide (oxide), silicon nitride, polycrystalline silicon (polysilicon or poly), and aluminum. A number of other materials can be deposited as thin films, including noble metals such as gold. However, noble metals will contaminate microelectronic circuitry causing it to fail, so any silicon wafers with noble metals on them have to be processed using equipment specially set aside for the purpose. Noble metal films are often patterned by a method known as "lift off," rather than wet or dry etching.

Often, photoresist is not tough enough to withstand the etching required. In such cases, a thin film of a tougher material (e.g. oxide or nitride) is deposited and patterned using photolithography. The oxide/nitride then acts as an etch mask during the etching of the underlying material. When the underlying material has been fully etched the masking layer is stripped away.

18.3.2.3 Wet Etching

Wet etching is a blanket name that covers the removal of material by immersing the wafer in a liquid bath of the chemical etchant. Wet etchants fall into two broad categories: isotropic etchants and anisotropic etchants. Isotropic etchants attack the material being etched at the same rate in all directions. Anisotropic etchants attack the silicon wafer at different rates in different directions, and so there is more control of the shapes produced. Some etchants attack silicon at different rates depending on the concentration of the impurities in the silicon (concentration-dependent etching).

Isotropic etchants are available for oxide, nitride, aluminum, polysilicon, gold, and silicon. Since isotropic etchants attack the material at the same rate in all directions, they remove material horizontally under the etch mask (undercutting) at the same rate as they etch through the material. This is illustrated for a thin film of oxide on a silicon wafer in Fig. 18.10, using an etchant that etches the oxide faster than the underlying silicon (e.g. hydrofluoric acid).

Anisotropic etchants are available to etch different crystal planes in silicon at different rates. The most popular anisotropic etchant is potassium hydroxide (KOH), since it is the safest to use.

Etching is done on a silicon wafer. Silicon wafers are slices that have been cut from a large ingot of silicon that was grown from a single seed crystal. The silicon atoms are all arranged in a crystalline structure, so the wafer is monocrystalline silicon (as opposed to polycrystalline silicon mentioned above). When purchasing

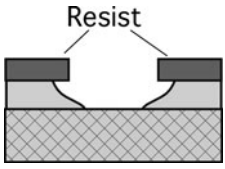

Fig. 18.10 Isotropic etching under the mask

silicon wafers, it is possible to specify that they have been sliced with the surface parallel to a particular crystal plane.

The simplest structures that can be formed using KOH to etch a silicon wafer with the most common crystal orientation (100) are shown in Fig. 18.11. These are the V-shaped groves, or pits with right-angled corners and sloping side walls. Using wafers with different crystal orientations can produce grooves or pits with vertical walls.

Both oxide and nitride etch slowly in KOH. Oxide can be used as an etch mask for short periods in the KOH etch bath (i.e. for shallow grooves and pits). For long periods, nitride is a better etch mask as it etches more slowly in the KOH.

KOH can also be used to produce mesa structures (Fig. 18.12a). When etching mesa structures, the corners can become beveled (Fig. 18.12b), rather than right angle corners. This has to be compensated for in some way. Typically, the etch mask is designed to include additional structures on the corners. These compensation structures are designed so that they are etched away entirely when the mesa is formed to leave 90° corners. One problem with using compensation structures to from right-angled mesa corners is that they put a limit on the minimum spacing between the mesas.

Fabrication of a diaphragm is one of the most popular sensor processes. It is used to produce accelerometers, pressures sensor, infrared temperature sensors (thermopiles and bolometers), and many others. Silicon diaphragms from about 50 μm thick

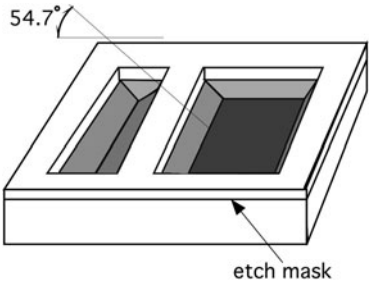

Fig. 18.11 Simple structures etched by KOH

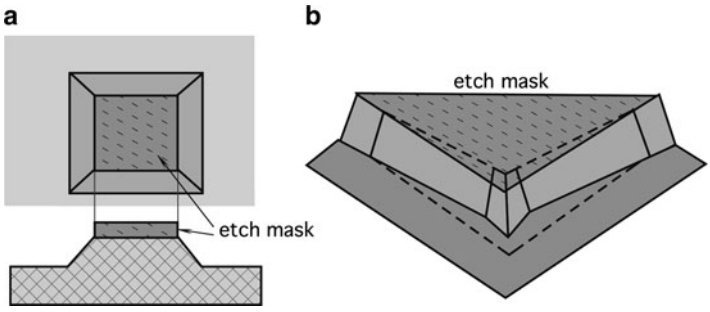

Fig. 18.12 Mesa structures

upwards can be made by etching through an entire wafer with KOH (Fig. 18.13a). The thickness is controlled by timing the etch, and so is subject to errors.

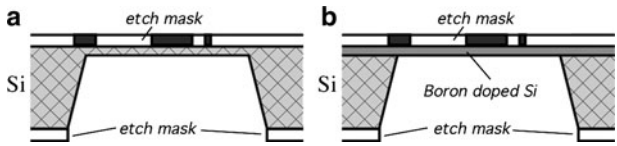

Fig. 18.13 Micromachining of a diaphragm or membrane

18.3.2.4 Concentration-Dependent Etching

Thinner diaphragms, of up to about 20 μm thick, can be produced using boron to stop the KOH etch (Fig. 18.13b). This is called the concentration-dependent etching. The thickness of the diaphragm is dependent on the depth to which the boron is diffused into the silicon, which can be controlled more accurately than the simple timed KOH etch. High levels of boron in silicon will reduce the rate at which it is etched in KOH by several orders of magnitude, effectively stopping the etching of the boron rich silicon. The boron impurities are usually introduced into the silicon by a process known as diffusion.

Besides the diaphragms, many other structures can be built by the concentration-dependent etching. A thick oxide mask is formed over the silicon wafer and patterned to expose the surface of the silicon wafer where the boron is to be introduced (Fig. 18.14a). The wafer is then placed in a furnace in contact with a boron diffusion source. Over a period of time, boron atoms migrate into the silicon wafer. Once the boron diffusion is completed, the oxide mask is stripped off (Fig. 18.14b). A second mask may then be deposited and patterned (Fig. 18.14c)

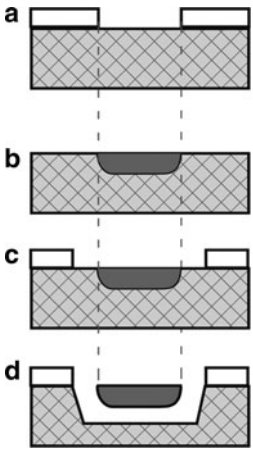

Fig. 18.14 Etching around the boron-doped silicon

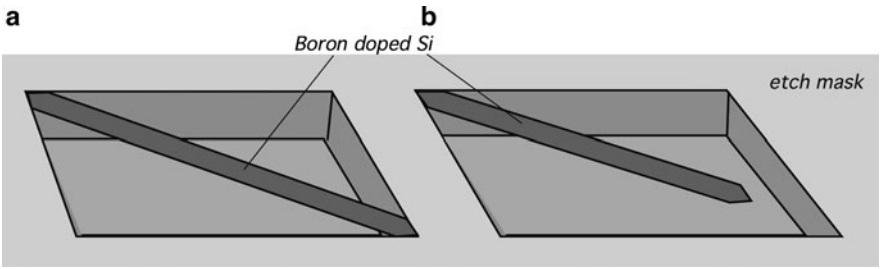

Fig. 18.15 Etching of a bridge and cantilever

before the wafer is immersed in the KOH etch bath. The KOH etches the silicon that is not protected by the mask, and etches around the boron-doped silicon (Fig. 18.14d). Boron can be driven into the silicon as far as 20 μm over periods of 15–20 h; however, it is desirable to keep the time in the furnace as short as possible. Concentration-dependent etching can also be used to produce narrow bridges or cantilever beams. Figure 18.15a shows a bridge, defined by a boron diffusion, spanning a pit that was etched from the front of the wafer in KOH. A cantilever beam (a bridge with one end free) produced by the same method is shown in Fig. 18.15b. The bridge and beam project across the diagonal of the pit to ensure that they will be etched free by the KOH. More complex structures are possible using this technique, but care must be taken to ensure that they will be etched free by the KOH.

One of the applications for these beams and bridges is the resonant sensors. The structure can be set vibrating at its fundamental frequency. Anything causing a change in the mass, length, etc. of the structure will register as a change frequency. Care has to be taken to ensure that only the quantity to be measured causes a significant change in frequency.

18.3.2.5 Dry Etching

The most common form of dry etching for micromachining applications is reactive ion etching (RIE). Ions are accelerated toward the material to be etched, and the etching reaction is enhanced in the direction of travel of the ion, thus RIE is an anisotropic etching technique. Deep trenches and pits (up to ten or a few tens of microns) of arbitrary shape and with vertical walls can be etched in a variety of materials including silicon, oxide, and nitride. Unlike anisotropic wet etching, RIE is not limited by the crystal planes in the silicon. A combination of dry etching and isotropic wet etching can be used to form very sharp points. First, a column with vertical sides is etched away using an RIE (Fig. 18.16a). A wet etch is then used, which undercuts the etch mask leaving a very fine point (Fig. 18.16b), the etch mask is then removed. Very fine points like this can be fabricated on the end of cantilever beams as probes for use, for example, in tactile sensors.

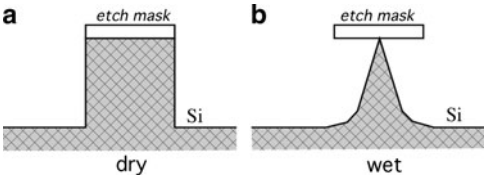

Fig. 18.16 Dry etching of a pointed structure

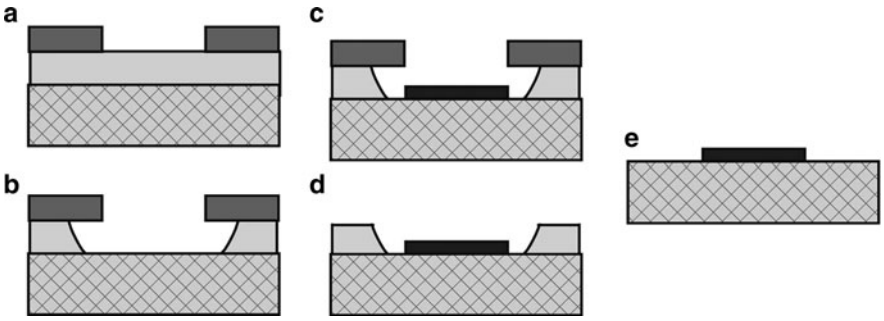

Fig. 18.17 Lift-off technique

18.3.2.6 Lift Off

Lift off is a stenciling technique often used to pattern noble metal films. There are a number of different techniques, the one outlined here is an assisted lift off method. A thin film of the assisting material (e.g. oxide) is deposited. A layer of resist is put over this and patterned, as for photolithography, to expose the oxide in the pattern desired for the metal (Fig. 18.17a). The oxide is then wet etched so as to undercut the resist (Fig. 18.17b). The metal is then deposited on the wafer, typically by a process known as evaporation (Fig. 18.17c). The metal pattern is effectively stenciled through the gaps in the resist, which is then removed lifting off the unwanted metal with it (Fig. 18.17d). The assisting layer is then stripped off too, leaving the metal pattern alone (Fig. 18.17e).

18.3.2.7 Wafer Bonding

There are a number of different methods available for bonding micromachined silicon wafers together, or to other substrates, to form larger more complex devices. A method of bonding silicon to glass that appears to be gaining in popularity is anodic bonding (electrostatic bonding). The silicon wafer and glass substrate are brought together and heated to a high temperature. A large electric field is applied across the join, which causes an extremely strong bond to form between the two materials. Figure 18.18 shows a glass plate bonded over a channel etched into a silicon wafer (RIE).

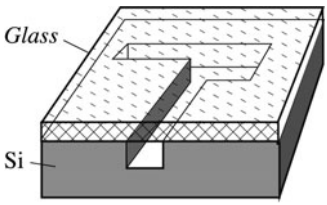

Fig. 18.18 Bonding of glass to silicon

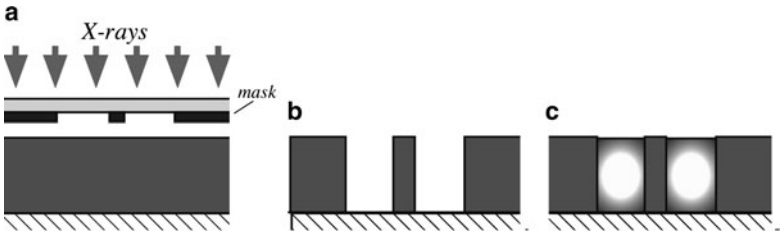

Fig. 18.19 LIGA technique to produce metal structure

It is also possible to bond silicon wafers directly together using gentle pressure, under water (direct silicon bonding). Other bonding methods include using an adhesive layer, such as a glass, or photoresist. While anodic bonding and direct silicon bonding form very strong joins, they suffer from some disadvantages, including the requirement that the surfaces to be joined are very flat and clean. Wafer bonding techniques can potentially be combined with some of the basic micromachined structures to form the membranes, cantilevers, valves, pumps, etc. of a microfluid-handling system that may be parts of chemical sensors.

18.3.2.8 LIGA

LIGA is capable of creating very finely defined microstructures of up to 1,000 μm high. In the process as originally developed, a special kind of photolithography using X-rays (X-ray lithography) is used to produce patterns in very thick layers of photoresist. The X-rays from a synchrotron source are shone through a special mask onto a thick photoresist layer (sensitive to X-rays), which covers a conductive substrate (Fig. 18.19a). This resist is then developed (Fig. 18.19b). The pattern formed is then electroplated with metal (Fig. 18.19c). The metal structures produced can be the final product; however, the metal structure can be used as a micromold that can be filled with various materials such as a plastic to produce the finished structures in that material. As the synchrotron source makes LIGA expensive, alternatives are being developed. These include high-voltage electron beam lithography, which can be used to produce structures of the order of 100 μm high, and excimer lasers capable of producing structures of up to several hundred microns high.

Naturally, electroplating is not limited to use with the LIGA process but may be combined with other processes and more conventional photolithography to produce microstructures.

References

1. Middelhoek S, Hoogerwerf AC (1985) Smart sensors: when and where? *Sens Actuators* 8(1):39–48 Elsevier Sequoya
2. Obermier E, Kopystynski P, Neißl R (1986) Characteristics of polysilicon layers and their application in sensors. In: *IEEE Solid-State Sensors Workshop*
3. Frijlink PM, Nicolas JL, Suchet P (1991) Layer uniformity in a multiwafer MOVPRE reactor for III–V compounds. *J Cryst Growth* 107:167–174
4. Morgan DV, Board K (1985) *An introduction to semiconductor microtechnology*, Wiley, New York
5. Muller RS, Howe RT, Senturia SD, Smith RL, White RM (eds) (1991) *Microsensors*, IEEE Press, Washington, DC
6. Rancourt JD (1996) *Optical thin films – user's handbook*, McGraw-Hill, New York
7. Ratner M, Ratner D(2003) *Nanotechnology: a gentle introduction to the next big idea*. Pearson Education, New York
8. Gad-El-Hak M (ed) (2006) MEMS – *introduction and fundamentals*. 2nd ed, CRC Press, Boca Raton, FL

Appendix

Table A.1 Chemical symbols for the elements

Ac	Actinium	Co	Cobalt	In	Indium	Os	Osmium	Sm	Samarium
Ag	Silver	Cr	Chromium	Ir	Iridium	P	Phosphorous	Sn	Tin
Al	Aluminum	Cs	Cesium	K	Potassium	Pa	Protactinium	Sr	Strontium
Am	Americium	Cu	Copper	Kr	Krypton	Pb	Lead	Ta	Tantalum
Ar	Argon	Dy	Dysprosium	La	Lanthanum	Pd	Palladium	Tb	Terbium
As	Arsenic	Er	Erbium	Li	Lithium	Pm	Promethium	Tc	Technetium
At	Astatine	Es	Einsteinium	Lr	Lawrencium	Po	Polonium	Te	Tellurium
Au	Gold	Eu	Europium	Lu	Lutetium	Pr	Praseodymium	Th	Thorium
B	Boron	F	Fluorine	Md	Mendelevium	Pt	Platinum	Ti	Titanium
Ba	Barium	Fe	Iron	Mg	Magnesium	Pu	Plutonium	Tl	Thallium
Be	Beryllium	Fm	Fermium	Mn	Manganese	Ra	Radium	Tm	Thulium
Bi	Bismuth	Fr	Francium	Mo	Molybdenum	Rb	Rubidium	U	Uranium
Bk	Berkelium	Ga	Gallium	N	Nitrogen	Re	Rhenium	V	Vanadium
Br	Bromine	Gd	Gadolinium	Na	Sodium	Rh	Rhodium	W	Tungsten
C	Carbon	Ge	Germanium	Nb	Niobium	Rn	Radon	Xe	Xenon
Ca	Calcium	H	Hydrogen	Nd	Neodymium	Ru	Ruthenium	Y	Yttrium
Cd	Cadmium	He	Helium	Ne	Neon	S	Sulfur	Yb	Ytterbium
Ce	Cerium	Hf	Hafnium	Ni	Nickel	Sb	Antimony	Zn	Zinc
Cf	Californium	Hg	Mercury	No	Nobelium	Sc	Scandium	Zr	Zirconium
Cl	Chlorine	Ho	Holmium	Np	Neptunium	Se	Selenium		
Cm	Curium	I	Iodine	O	Oxygen	Si	Silicon		

Table A.2 SI multiples

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	hecto	h	10^{-15}	femto	f
10^{1}	deka	da	10^{-18}	atto	a

Table A.3 Derivative SI units

Quantity	Name of unit	Expression in terms of basic units
Area	square meter	m^2
Volume	cubic meter	m^3
Frequency	hertz (Hz)	s^{-1}
Density (concentration)	kilogram per cubic meter	kg/m^3
Velocity	meter per second	m/s
Angular velocity	radian per second	rad/s
Acceleration	meter per second squared	m/s^2
Angular acceleration	radian per second squared	rad/s^2
Volumetric flow rate	cubic meter per second	m^3/s
Force	newton (N)	$kg\ m/s^2$
Pressure	newton per square meter (N/m^2) or pascal (Pa)	$kg/m\ s^2$
Work energy heat torque	joule (J), newton-meter (N m) or watt-second (W s)	$kg\ m^2/s^2$
Power heat flux	watt (W) Joule per second (J/s)	$kg\ m^2/s^3$
Heat flux density	watt per square meter (W/m^2)	kg/s^3
Specific heat	joule per kilogram degree (J/kg deg)	m^2/s^2 deg
Thermal conductivity	watt per meter degree (W/m deg) or (J m/s m^2 deg)	$kg\ m/s^3$ deg
Mass flow rate (mass flux)	kilogram per second	kg/s
Mass flux density	kilogram per square meter-second	kg/m^2 s
Electric charge	coulomb (C)	A s
Electromotive force	volt (V) or (W/A)	$kg\ m^2/A\ s^3$
Electric resistance	ohm (Ω) or (V/A)	$kg\ m^2/A^2\ s^3$
Electric conductivity	ampere per volt-meter (A/V m)	$A^2s^3/kg\ m^3$
Electric capacitance	farad (F) or (A s/V)	$A^3\ s^4/kg\ m^2$
Magnetic flux	weber (Wb) or (V s)	$kg\ m^2/A\ s^2$
Inductance	henry (H) or (Vs/A)	$kg\ m^2/A^2\ s^2$
Magnetic permeability	henry per meter (H/m)	$kg\ m/A^2\ s^2$
Magnetic flux density	tesla (T) or weber per square meter (Wb/m^2)	$kg/A\ s^2$
Magnetic field strength	ampere per meter	A/m
Magnetomotive force	ampere	A
Luminous flux	lumen (lm)	cd sr
Luminance	candela per square meter	cd/m^2
Illumination	lux (lx) or lumen per square meter (lm/m^2)	$cd\ sr/m^2$

Tables A.4 SI conversion multiples (to make a conversion to SI, a non-SI value should be multiplied by a number given in the table)

Acceleration: (m/s^2)			
Ft/s^2	0.3048	gal	0.01
free fall (g)	9.80665	in/s^2	0.0254
Angle: radian (rad)			
degree	0.01745329	second	4.848137×10^{-6}
minute	2.908882×10^{-4}	grade	1.570796×10^{-2}
Area: (m^2)			
acre	4,046.873	hectare	1×10^4
are	100.00	mi^2 (US statute)	2.589998×10^6
ft^2	9.290304×10^{-2}	yd^2	0.8361274

(*continued*)

Tables A.4 (continued)

Bending moment or torque: (N m)			
dyne cm	1×10^{-7}	lbf in	0.1129848
kgf m	9.806650	lbf ft	1.355818
ozf in	7.061552×10^{-3}		
Electricity and Magnetism[a]			
ampere hour	3,600 coulomb (C)	EMU of inductance	8.987×10^{11} henry (H)
EMU of capacitance	10^{9} farad (F)	EMU of resistance	8.987×10^{11} (Ω)
EMU of current	10 ampere (A)	faraday	9.65×10^{19} coulomb (C)
EMU of elec. potential	10^{-8} volt (V)	gamma	10^{-9} tesla (T)
EMU of inductance	10^{-9} henry (H)	gauss	10^{-4} tesla (T)
EMU of resistance	10^{-9} ohm (Ω)	gilbert	0.7957 ampere (A)
ESU of capacitance	1.112×10^{-12} farad (F)	maxwell	10^{-8} weber (Wb)
ESU of current	3.336×10^{-10} ampere (A)	mho	1.0 siemens (S)
EMU of elec. potential	299.79 volt (V)	ohm centimeter	0.01 ohm meter (Ω m)
Energy (work): joule (J)			
British thermal unit (Btu)	1,055	kilocalorie	4,187
calorie	4.18	kW h	3.6×10^{6}
calorie (kilogram)	4,184	ton (nuclear equiv. TNT)	4.184×10^{9}
electronvolt	1.60219×10^{-19}	therm	1.055×10^{8}
erg	10^{-7}	W h	3,600
ft lbf	1.355818	W s	1.0
ft-poundal	0.04214		
Force**:** newton (N)			
dyne	10^{-5}	ounce-force	0.278
kilogram-force	9.806	pound-force (lbf)	4.448
kilopond (kp)	9.806	poundal	0.1382
kip (1,000 lbf)	4,448	ton-force (2,000 lbf)	8,896
Heat			
Btu ft/(h ft2 °F) (thermal conductivity)	1.7307 W/(m K)	cal/cm^2	4.18×10^{4} J/m^2
Btu/lb	2,324 J/kg	cal/(cm^2 min)	697.3 W/m^2
Btu/(lb °F) = cal/(g °C) (heat capacity)	4,186 J/(kg K)	cal/s	4.184 W
Btu/ft^3	3.725×10^{4} J/m^3	°F h ft^2/Btu (thermal resistance)	0.176 K m^2/W
cal/(cm s °C)	418.4 W/(m K)	ft^2/h (thermal diffusivity)	2.58×10^{-5} m^2/s
Length: meter (m)			
angstrom	10^{-10}	microinch	2.54×10^{-8}
astronomical unit	1.495979×10^{11}	micrometer (micron)	10^{-6}
chain	20.11	mil	2.54×10^{-5}
fermi (femtometer)	10^{-15}	mile (nautical)	1,852.000
foot	0.3048	mile (international)	1,609.344

(*continued*)

Tables A.4 (continued)

inch	0.0254	pica (printer's)	4.217×10^{-3}
light year	9.46055×10^{15}	yard	0.9144
Light			
cd/in^2	1,550 cd/m^2	lambert	3.183×10^3 cd/m^2
footcandle	10.76 lx (lux)	lm/ft^2	10.76 lm/m^2
footlambert	3.426 cd/m^2		
Mass**:** kilogram (kg)			
carat (metric)	2×10^{-4}	ounce (troy or apothecary)	3.110348×10^{-2}
grain	6.479891×10^{-5}	pennyweight	1.555×10^{-3}
gram	0.001	pound (lb avoirdupois)	0.4535924
hundredweight (long)	50.802	pound (troy or apothecary)	0.3732
hundredweight (short)	45.359	slug	14.5939
kgf s^2/m	9.806650	ton (long, 2,240 lb)	907.184
ounce (avoirdupois)	2.834952×10^{-2}	ton (metric)	1,000
Mass: per unit time (includes Flow)			
perm (0°C)	5.721×10^{-11} kg/(Pa s m^2)	lb/(hp h) SPC – specific fuel consumption	1.689659×10^{-7} kg/J
lb/h	1.2599×10^{-4} kg/s	ton (short)/h	0.25199 kg/s
lb/s	0.4535924 kg/s		
Mass per unit volume (includes Density and Capacity): (kg/m^3)			
oz (avoirdupois)/gal (UK liquid)	6.236	oz (avoirdupois)/gal (US liquid)	7.489
oz (avoirdupois)/in^3	1,729.99	$slug/ft^3$	515.3788
lb/gal (US liquid)	11.9826 kg/m^3	ton (long)/yd^3	1,328.939
Power**:** watt (W)			
Btu (International)/s	1,055.056	horsepower (electric)	746
cal/s	4.184	horsepower (metric)	735.499
erg/s	10^{-7}	horsepower (UK)	745.7
horsepower (550 ft lbf/s)	745.6999	ton of refrigeration (12,000 Btu/h)	3,517
Pressure or Stress: pascal (Pa)			
atmosphere, standard	1.01325×10^5	$dyne/cm^2$	0.1
atmosphere, technical	9.80665×10^4	foot of water (39.2°F)	2,988.98
bar	10^5	$poundal/ft^2$	1.488164
centimeter of mercury (0°C)	1,333.22	psi (lbf/in^2)	6,894.757
centimeter of water (4°C)	98.0638	torr (mmHg, 0°C)	133.322
Radiation units			
curie	3.7×10^{10} becquerel (Bq)	rem	0.01 sievert (Sv)
rad	0.01 gray (Gy)	roentgen	2.58×10^{-4} C/kg
Temperature			
°Celsius	TK = t°C + 273.15 K	°Fahrenheit	T°C = (t°F − 32)/1.8°C
°Fahrenheit	TK = (t°F + 459.67)/1.8 K	°Rankine	TK = T°R/1.8

(*continued*)

Tables A.4 (continued)

Velocity (includes Speed): (m/s)			
ft/s	0.3048	mi/h (International)	0.44704
in/s	2.54×10^{-2}	rpm (r/min)	0.1047 rad/s
knot (International)	0.51444		
Viscosity: (Pa s)			
centipose (dynamic viscosity)	10^{-3}	lbf s/in^2	6,894.757
centistokes (kinematic viscosity)	10^{-6}	rhe	10 1/(Pa s)
poise	0.1	slug/(ft s)	47.88026
poundal s/ft^2	1.488164	stokes	10^{-4} m^2/s
lb/(ft s)	1.488164		
Volume (includes Capacity): (m^3)			
acre-foot	1,233.489	gill (U.S.)	1.182941×10^{-4}
barrel (oil, 42 gal)	0.1589873	in^3	1.638706×10^{-5}
bushel (U.S.)	3.5239×10^{-2}	liter	10^{-3}
cup	2.36588×10^{-4}	ounce (U.S. fluid)	2.957353×10^{-5}
ounce (U.S. fluid)	2.95735×10^{-5}	pint (U.S. dry)	5.506105×10^{-4}
ft^3	2.83168×10^{-2}	pint (U.S. liquid)	4.731765×10^{-4}
gallon (Canadian, U.K. liquid)	4.54609×10^{-3}	tablespoon	1.478×10^{-5}
gallon (U.S. liquid)	3.7854×10^{-3}	ton (register)	2.831658
gallon (U.S. dry)	4.40488×10^{-3}	yd^3	0.76455

[a]ESU means electrostatic cgs unit; EMU means electromagnetic cgs unit

Table A.5 Dielectric constants of some materials at room temperature (25°C)

Material	κ	Frequency (Hz)	Material	κ	Frequency (Hz)
Air	1.00054	0	Paraffin	2.0–2.5	10^6
Alumina ceramic	8–10	10^4	Plexiglas	3.12	10^3
Acrylics	2.5–2.9	10^4	Polyether sulfone	3.5	10^4
ABS/Polysulfone	3.1	10^4	Polyesters	3.22–4.3	10^3
Asphalt	2.68	10^6	Polyethylene	2.26	10^3–10^8
Beeswax	2.9	10^6	Polypropylenes	2–3.2	10^4
Benzene	2.28	0	Polyvinyl chloride	4.55	10^3
Carbon tetrachloride	2.23	0	Porcelain	6.5	0
Cellulose nitrate	8.4	10^3	Pyrex glass (7070)	4.0	10^6
Ceramic (titanium dioxide)	14–110	10^6	Pyrex glass (7760)	4.5	0
Cordierite	4–6.23	10^4	Rubber (neoprene)	6.6	10^3
Compound for thick film capacitors	300–5,000	0	Rubber (silicone)	3.2	10^3
Diamond	5.5	10^8	Rutile $\perp$ optic axis	86	10^8
Epoxy resins	2.8–5.2	10^4	Rutile $\parallel$ optic axis	170	10^8
Ferrous oxide	14.2	10^8	Silicone resins	3.4–4.3	10^4
Flesh (skin, blood, muscles)	97	40×10^6	Tallium chloride	46.9	10^8
Flesh (fat, bones)	15	40×10^6	Teflon	2.04	10^3–10^8
Lead nitrate	37.7	6×10^7	Transformer oil	4.5	0
Methanol	32.63	0	Vacuum	1	–
Nylon	3.5–5.4	10^3	Water	78.5	0
Paper	3.5	0			

Table A.6 Properties of magnetic materials (adapted from Sprague, CN-207 Hall effect IC applications, 1986)

Material	MEP (G Oe) 10^6	Residual induction (G) 10^3	Coercive force (Oe) 10^3	Temperature coefficient %/°C	Cost
R.E.-Cobalt	16	8.1	7.9	−0.05	Highest
Alnico 1,2,3,4	1.3–1.7	5.5–7.5	0.42–0.72	−0.02 to −0.03	Medium
Alnico 5,6,7	4.0–7.5	10.5–13.5	0.64–0.78	−0.02 to −0.03	Medium/high
Alnico 8	5.0–6.0	7–9.2	1.5–1.9	−0.01 to 0.01	Medium/high
Alnico 9	10	10.5	1.6	−0.02	high
Ceramic 1	1.0	2.2	1.8	−0.2	low
Ceramic 2,3,4,6	1.8–2.6	2.9–3.3	2.3–2.8	−0.2	low/medium
Ceramic 5,7,8	2.8–3.5	3.5–3.8	2.5–3.3	−0.2	Medium
Cunife	1.4	5.5	0.53	–	Medium
Fe-Cr	5.25	13.5	0.6	–	Medium/high
Plastic	0.2–1.2	1.4	0.45–1.4	−0.2	Lowest
Rubber	0.35–1.1	1.3–2.3	1–1.8	−0.2	Lowest

Table A.7 Resistivities (ρ) $10^{-8}\Omega\cdot$m ((at room temperature) and temperature coefficients of resistivity (TCR) 10^{-3}/°K of some materials at room temperature)

Material	ρ	TCR (α)	Material	ρ	TCR (α)
Alumina[a]	$>10^{20}$		Palladium	10.54	3.7
Aluminum (99.99%)	2.65	3.9	Platinum	10.42	3.7
Beryllium	4.0	0.025	Platinum + 10% Rhodium	18.2	
Bismuth	10^6		Polycrystalline glass[a]	6.3×10^{14}	
Brass (70Cu, 30Zn)	7.2	2.0	Rare earth metals	28–300	
Carbon	3,500	–0.5	Silicon (very sensitive to purity)	$(3.4 \text{ to } 15)\times10^6$	
Chromium plating	14–66		Silicon bronze (96Cu, 3Si, 1Zn)	21.0	
Constantan (60Cu, 40Ni)	52.5	0.01	Silicon nitride	10^{19}	
Copper	1.678	3.9	Silver	1.6	6.1
Evanohm (75Ni, 20Cr, 2.5Al, 2.5Cu)	134		Sodium	4.75	
Germanium (polycrystalline)	46×10^6		Stainless steel (cast)	70–122	
Gold	2.24	3.4	Tantalum	12.45	3.8
Iridium	5.3		Tantalum carbide	20	
Iron (99.99%)	9.71	6.5	Tin	11.0	4.7
Lead	22	3.36	Titanium	42	
Manganese	185		Titanium and its alloys	48–199	
Manganin	44	0.01	Titanium carbides	105	
Manganin (84Cu, 12Mn, 4Ni)	48		Tungsten	5.6	4.5
Mercury	96	0.89	Zinc	5.9	4.2
Mullite[a]	10^{21}		Zircon[a]	$>10^{20}$	
Nichrome	100	0.4	Zirconium and its alloys	40–74	
Nickel	6.8	6.9			

[a]Volume resistivity

Table A.8 Properties of piezoelectric materials at 20°C

	PVDF	$BaTiO_3$	PZT	Quartz	TGS
Density ($\times 10^3$ kg/m^3)	1.78	5.7	7.5	2.65	1.69
Dielectric constant, ε_r	12	1,700	1,200	4.5	45
Elastic modulus (10^{10} N/m)	0.3	11	8.3	7.7	3
Piezoelectric constant (pC/N)	$d_{31} = 20$ $d_{32} = 2$ $d_{33} = -30$	78	110	2.3	25
Pyroelectric constant (10^{-4} C/m^{2}K)	4	20	27	–	30
Electromechanical coupling constant (%)	11	21	30	10	–
Acoustic impedance (10^6 kg/m^{2}s)	2.3	25	25	14.3	–

Table A.9 Physical properties of pyroelectric materials (from Meixner H, Mader G, and Kleinschmidt P (1986) Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). Siemens Forsch-u Entwicl Ber Bd 15(3):105–114)

Material	Curie temperature (°C)	Thermal conductivity (W/(mK))	Relative permitivity (ε_r)	Pyroelectric charge coef. (C/(m^{2}K))	Pyroelectric voltage coef. (V/(mK))	Coupling ($k_p{}^2$ (%))
Single crystals						
TGS	49	0.4	30	3.5×10^{-4}	1.3×10^6	7.5
$LiTa0_3$	618	4.2	45	2.0×10^{-4}	0.5×10^6	1.0
Ceramics						
$BaTiO_3$	120	3.0	1,000	4.0×10^{-4}	0.05×10^6	0.2
PZT	340	1.2	1,600	4.2×10^{-4}	0.03×10^6	0.14
Polymers						
PVDF	205	0.13	12	0.4×10^{-4}	0.40×10^6	0.2
Polycrystalline layers						
$PbTiO_3$	470	2 (monocrystal)	200	2.3×10^{-4}	0.13×10^6	0.39

Note: The above figures may vary depending on manufacturing technologies

Table A.10 Characteristics of thermocouple types

Junction materials	Sensitivity μV/°C (@ 25°C)	Temperature range (°C)	Applications	Designation
Copper/Constantan	40.9	−270 to +600	Oxidation, reducing, inert, vacuum. Preferred below 0°C. Moisture resistant	T
Iron/Constantan	51.7	−270 to +1,000	Reducing and inert atmosphere. Avoid oxidation and moisture	J
Chromel/Alumel	40.6	−270 to 1,300	Oxidation and inert atmospheres	K
Chromel/Constantan	60.9	−200 to 1,000		E

(*continued*)

Table A.10 (continued)

Junction materials	Sensitivity μV/°C (@ 25°C)	Temperature range (°C)	Applications	Designation
Pt (10%)/Rh-Pt	6.0	0–1,550	Oxidation and inert atmospheres, avoid reducing atmosphere and metallic vapors	S
Pt (13%)/Rh-Pt	6.0	0–1,600	Oxidation and inert atmospheres, avoid reducing atmosphere and metallic vapors	R
Slver-Paladium	10.0	200–600		
Constantan-Tungsten	42.1	0–800		
Silicon-Aluminum	446	−40 to 150	Used in thermopiles and micromachined sensors	

Table A.11 Thermoelectric coefficients and volume resistivities of selected elements (adapted from Schieferdecker J et al (1995) Infrared thermopile sensors with high sensitivity and very low temperature coefficient. Sens Actuators A 46–47:422–427)

Element	α (μV K^{-1})	ρ (μΩ m)
p-Si	100–1,000	10–500
p-Poly-Si	100–500	10–1,000
Antimony (Sb)	32	18.5
Iron (Fe)	13.4	0.086
Gold (Au)	0.1	0.023
Copper (Cu)	0	0.0172
Silver (Ag)	−0.2	0.016
Aluminum (Al)	−3.2	0.028
Platinum (Pt)	−5.9	0.0981
Cobalt (Co)	−20.1	0.0557
Nickel (Ni)	−20.4	0.0614
Bismuth (Bi)	−72.8	1.1
n-Si	−100 to −1,000	10–500
n-Poly-Si	−100 to −500	10–1,000

Table A.11a Thermocouples for very low and very high temperatures

Materials	Useful range °C	Approx sensitivity μV/°C
Iron-Constantan	down to −272	−32
Copper-Constantan	down to −273	−22.9
Cromel-Alumel	down to –272	−23.8
Tantalum-Tungsten	up to 3,000	6.1
Tangsten-Tangsten(50)Molybdenum	up to 2,900	2.8
Tangsten-Tangsten(20)Rhenium	up to 2,900	12.7

Table A.12 Densities (kg/m^3) of some materials at 1atm pressure and 0°C

Best laboratory vacuum	10^{-17}	Silica	1,938–2,657
Hydrogen	0.0899	Graphite recrystallized	1,938
Helium	0.1785	Borosilicate glass (TEMPAX®)[a]	2,200
Methane	0.7168	Asbestos fibers	2,400–3,300
Carbon monoxide	1.250	Silicon	2,333
Air	1.2928	Polycrystalline glass	2,518–2,600
Oxygen	1.4290	Aluminum	2,700
Carbon dioxide	1.9768	Mullite	2,989–3,293
Plastic foams	10–600	Silicon nitride	3,183
Benzene	680–740	Alumina ceramic	3,322–3,875
Alcohol	789.5	Zinc alloys	5,200–7,170
Turpentine	860	Vanadium	6,117
Mineral oil	900–930	Chromium	7,169
Natural rubber	913	Tin and its alloys	7,252–8,000
Polyethylene, low density	913	Stainless steel	8,138
Ice	920	Bronzes	8,885
Polyethylene, high density	950	Copper	8,941
Carbon and graphite fibers	996–2,000	Cobalt and its alloys	9,217
Water	1,000	Nickel and its alloys	9,245
Nylon 6	1,100	Bismuth	9,799
Hydrochloric acid (20%)	1,100	Silver	10,491
Acrylics	1,163–1,190	Lead and its alloys	11,349
Epoxies	1,135–2,187	Palladium	12,013
Coal tar	1,200	Mercury	13,596
Phenolic	1,246–2,989	Molybdenum	13,729
Glycerin	1,260	Tantalum and its alloys	16,968
PVC	1,350	Gold	19,320
Saran fibers	1,700	Tungsten and its alloys	19,653
Sulfuric acid (20%)	1,700	Platinum	21,452
Polyester	1,800	Iridium	22,504
Beryllium and its alloys	1,855–2,076	Osmium	22,697

[a]TEMPAX® is a registered trademark of Schott Glasswerke, Mainz, Germany

Table A.13 Mechanical properties of some solid materials

Material	Modulus of elasticity (GPa)	Poisson's ratio (ν)	density (kg/m^3)
Aluminum	71	0.334	2,700
Beryllium copper	124	0.285	8,220
Brass	106	0.324	8,530
Copper	119	0.326	8,900
Glass	46.2	0.245	2,590
Lead	36.5	0.425	11,380
Molybdenum	331	0.307	10,200
Phosphor bronze	11	0.349	8,180
Steel (Carbon)	207	0.292	7,800
Steel (Stainless)	190	0.305	7,750

Table A.14 Mechanical properties of some crystalline materials (from Petersen KE (1982) Silicon as a mechanical material. Proc IEEE 70(5):420–457)

Material	Yield strength (10^{10} dyne/cm^2)	Knoop hardness (kg/mm^2)	Young's modulus (10^{12} dyne/cm^2)	Density (g/cm^3)	Thermal conductivity (W/cm°C)	Thermal expansion (10^{-6}/°C)
Diamond[a]	53	7,000	10.35	3.5	20.0	1.0
SiC[a]	21	2,480	7.0	3.2	3.5	3.3
TiC[a]	20	2,470	4.97	4.9	3.3	6.4
Al_2O_3[a]	15.4	2,100	5.3	4.0	0.5	5.4
Si_3N_4[a]	14	3,486	3.85	3.1	0.19	0.8
Iron[a]	12.6	400	1.96	7.8	0.803	12.0
SiO_2 (fibers)	8.4	820	0.73	2.5	0.014	0.55
Si[a]	7.0	850	1.9	2.3	1.57	2.33
Steel (max. strength)	4.2	1,500	2.1	7.9	0.97	12.0
W	4.0	485	4.1	19.3	1.78	4.5
Stainless Steel	2.1	660	2.0	7.9	0.329	17.3
Mo	2.1	275	3.43	10.3	1.38	5.0
Al	0.17	130	0.70	2.7	2.36	25.0

[a]Single crystal

Table A.15 Speed of sound waves

Medium	Speed (m/s)
Air (dry at 20°C)	331
Steam (134°C)	494
Hydrogen (20°C)	1,330
Water (fresh)	1,486
Water (sea)	1,519
Lead	1,190
Copper	3,810
Aluminum	6,320
Pyrex® glass	5,170
Steel	5,200
Beryllium	12,900

Gases at 1 atm pressure, solids in long thin rods

Table A.16 Coefficient (α) of linear thermal expansion of some materials (per °C$\times 10^{-6}$)

Material	α	Material	α
Alnico I (Permanent magnet)	12.6	Nylon	90
Alumina (polycrystalline)	8.0	Phosphor-bronze	9.3
Aluminum	25.0	Platinum	9.0
Brass	20.0	Plexiglas (Lucite)	72
Cadmium	30.0	Polycarbonate (ABS)	70
Chromium	6.0	Polyethylene (high density)	216
Comol (Permanent magnet)	9.3	Silicon	2.6
Copper	16.6	Silver	19.0

(*continued*)

Table A.16 (continued)

Material	α	Material	α
Fused quartz	0.27	Solder 50–50	23.6
Glass (Pyrex®)	3.2	Steel (SAE 1020)	12.0
Glass (regular)	9.0	Steel (stainless: type 304)	17.2
Gold	14.2	Teflon	99
Indium	18.0	Tin	13.0
Invar	0.7	Titanium	6.5
Iron	12.0	Tungsten	4.5
Lead	29.0	Zinc	35.0
Nickel	11.8		

Table A.17 Specific heat and thermal conductivity of some materials (at 25°C)

Material	Specific heat $\left(\frac{J}{kg\,^{\circ}C}\right)$	Thermal conductivity $\left(\frac{W}{m\,^{\circ}C}\right)$	Density $\left(\frac{kg}{m^3}\right)$
Air (1 atm)	995.8	0.024	1.2
Alumina	795	6	4,000
Aluminum	481	88–160	2,700
Bakelite	1,598	0.23	1,300
Brass	381	26–234	8,500
Chromium	460	91	
Constantan	397	22	8,800
Copper	385	401	8,900
Diamond		99–232	
Fiberglass	795	0.002–0.4	60
Germanium		60	
Glass (Pyrex)	780	0.1	2,200
Glass (regular)		1.9–3.4	
Gold	130	296	19,300
Graphite		112–160	
Iron	452	79	7,800
Lead	130	35	11,400
Manganin	410	21	8,500
Mercury	138	8.4	13,500
Nickel and its alloys	443	6–50	8,900
Nylon	1,700	0.24	1,100
Platinum	134	73	21,400
Polyester	1,172	0.57–0.73	1,300
Polyurethane foam		0.024	40
Silicon	668	83.7	2,333
Silicone oil	1,674	0.1	900
Silver	238	419	10,500
Stainless steel	460	14–36	8,020
Styrofoam	1,300	0.003–0.03	50
Teflon TFE	998	0.4	2,100
Tin	226	64	7,300
Tungsten	139	96.6	19,000
Water	4,184	0.6	1,000
Zinc	389	115–125	7,100

Table A.18 Typical emissivities of different materials (from 0 to 100°C)

Material	Emissivity	Material	Emissivity
Blackbody (ideal)	*1.00*	Green leaves	0.88
Cavity Radiator	*0.99–1.00*	Ice	0.96
Aluminum (anodized)	0.70	Iron or Steel (rusted)	0.70
Aluminum (oxidized)	0.11	Nickel (oxidized)	0.40
Aluminum (polished)	0.05	Nickel (unoxidized)	0.04
Aluminum (rough surface)	0.06–0.07	Nichrome (80Ni-20Cr) (oxidized)	0.97
Asbestos	0.96	Nichrome (80Ni-20Cr) (polished)	0.87
Brass (dull tarnished)	0.61	Oil	0.80
Brass (polished)	0.05	Silicon	0.64
Brick	0.90	Silicone Rubber	0.94
Bronze (polished)	0.10	Silver (polished)	0.02
Carbon filled latex paint	0.96	Skin (human)	0.93–0.96
Carbon Lamp Black	0.96	Snow	0.85
Chromium (polished)	0.10	Soil	0.90
Copper (oxidized)	0.6–0.7	Stainless Steel (buffed)	0.20
Copper (polished)	0.02	Steel (flat rough surface)	0.95–0.98
Cotton cloth	0.80	Steel (ground)	0.56
Epoxy Resin	0.95	Tin plate	0.10
Glass	0.95	Water	0.96
Gold	0.02	White Paper	0.92
Gold-black	0.98–0.99	Wood	0.93
Graphite	0.7–0.8	Zinc (polished)	0.04

Table A.19 Refractive indices (n) of some materials

Material	n	wavelength (μm)	note
Vacuum	1		
Air	1.00029		
Acrylic	1.5	0.41	
AMTIR-1	2.6	1	Amorphous glass[a]
($Ge_{33}As_{12}Se_{55}$)	2.5	10	
AMTIR-3	2.6	10	Amorphous glass[a]
($Ge_{28}Sb_{12}Se_{60}$)			
As_2S_3	2.4	8.0	Amorphous glass[a]
CdTe	2.67	10.6	
Crown glass	1.52		
Diamond	2.42	0.54	Excellent thermal conductivity
Fused silica (SiO_2)	1.46	3.5	
Borosilicate glass	1.47	0.7	TEMPAX®[b]. Transparent: 0.3–2.7μm
GaAs	3.13	10.6	Laser windows
Germanium	4.00	12.0	
Heaviest flint glass	1.89		
Heavy flint glass	1.65		
Irtran 2 (ZnS)	2.25	4.3	Windows in IR sensors
KBr	1.46	25.1	Hygroscopic
KCl	1.36	23.0	Hygroscopic
KRS-5	2.21	40.0	Toxic
KRS-6	2.1	24	Toxic
NaCl	1.89	0.185	Hygroscopic, corrosive
Polyethylene	1.54	8.0	Low cost IR windows/lenses

(*continued*)

Table A.19 (continued)

Material	n	wavelength (μm)	note
Polystyrene	1.55		
Pyrex 7740	1.47	0.589	Good thermal and optical properties
Quartz	1.54		
Sapphire (Al_2O_3)	1.59	5.58	Chemically resistant
Silicon	3.42	5.0	Windows in IR sensors
Silver Bromide (AgBr)	2.0	10.6	Corrosive
Silver Chloride (AgCl)	1.9	20.5	Corrosive
Water [20°C]	1.33		
ZnSe	2.4	10.6	IR windows, brittle

[a]Available from Amorphous Materials, Inc. Garland, TX 75042
[b]TEMPAX® is a registered trademark of Schott Glasswerke, Mainz, Germany

Table A.20 Characteristics of C-Zn and Alkaline cells (from Powers RA (1995) Batteries for low power electronics. Proc IEEE83(4):687–693)

Battery	Wh/L	Wh/kg	Drain rate	Shelf life
Carbon-Zinc	150	85	Low-medium	2 years
Alkaline	250	105	Medium-high	5 years

Table A.21 Lithium-manganese dioxide primary cells (from Powers RA (1995) Batteries for low power electronics. Proc IEEE 83(4):687–693)

Construction	Voltage	Capacity (mAh)	Rated d.c. current (mA)	Pulse current (mA)	Energy density (W h/L)
Coin	3	30–1,000	0.5–7	5–20	500
Cyl. Wound	3	160–1,300	20–1,200	80–5,000	500
Cyl. Bobbin	3	650–500	4–10	60–200	620
Cyl. "D" cell	3	10,000	2,500		575
Prismatic	3	1,150	18		490
Flat	3/6	150–1,400	20–125		290

Table A.22 Typical characteristics of "AA"-size secondary cells

System	Volts	Capacity (mAh)	Rate (C)[a]	W h/L	W h/kg	Cycles	Loss/Mo (%)
NiCad	1.2	1,000	10	150	60	1,000	15
Ni-MH	1.2	1,200	2	175	65	500	20
Pb Acid	2	400	1	80	40	200	2
Li Ion (CoO_2)	3.6	500	1	225	90	1,200	8
Li/MnO_2	3	800	0.5	280	130	200	1

[a]Discharge rate unit, C, (in mA) is equal numerically to the nominal capacity (in mA h).

Table A.23 Miniature secondary cells and batteries

Manufacturer	Part	Type	Size	Capacity (mAh)	Voltage	Price $ (appx)
Avex Corp. Bensalem, PA 800-345-1295		RAM	AA	1.4	1.5	1
GN National Electric Inc. Pomona, CA 909-598-1919	GN-360	NiCd	15.5×19 mm	60	3.6	1.10
GP Batteries USA, San Diego, CA 619–674–5620	Green-Charge	NiMH	2/3AA, AA, 2/3 AF, 4/5AF	600–2,500	1.2	2–7
Gould, Eastlake, OH 216–953–5084	3C120M	$LiMnO^2$	3×4×0.12 cm	120	3	2.71
House of Batteries Inc., Huntington Beach, CA 800–432–3385	Green Cell	NiMH	AA, 4/5A, 7/5A	1,200–2,500	1–2	3.50–12
Maxell Corp., Fairlawn, NJ 201–794–5938	MHR-AAA	NiMH	AAA	410	1.2	4
Moli Energy Ltd., Maple Ridge, BC, Canada, 604–465–7911	MOLICEL	Li-ion	18(dia)×65 mm	1,200	3.0–4.1	25
Plainview Batteries, Inc., Plainview, NY 516–249–2873	PH600	NiMH	48×17×7.7 mm	600	1.2	4
Power Coversion, Inc., Elmwood Park, NJ 201–796–4800	MO4/11	$LiMnO^2$	1/2AA	1,000	3.3	5–8
Power Sonic Corp., Redwood City, CA, 415–364–5001	PS-850AA	NiCd	AA	850	1.2	1.75
Rayovac Corp., Madison, WI 608–275–4690	Renewal	RAM	AA; AAA	1,200; 600	1.5	from 0.50
Renata U.S., Richardson, TX 214–234–8091	CR1025	Li	10 mm	25	3.0	0.50
Sanyo Energy (U.S.A.), San Diego, CA, 691–661–7992	Twicell	NiMH	10.4×44.5 ×67 mm	450	1.2	3.85
Saft America, Inc., San Diego, CA, 619–661–7992	VHAA	NiMH	AA	1,100	1.2	2.95
Tadiran Electronics, Port Washington, NY, 516–621–4980		Li	1/AA-DD packs	370 mAh to 30 Ah	3–36	1+

(*continued*)

Table A.23 (continued)

Manufacturer	Part	Type	Size	Capacity (mAh)	Voltage	Price $ (appx)
Toshiba America, Deerfield, IL, 800–879–4963	LSQ8	Li-ion	8.6×3.4 ×48 mm	900	3.7	12–15
Ultralife Batteries, Inc., Newark, NJ, 315–332–7100	U3VL	Li	25.8×44.8×16.8	3,600	3.0	4.60
Varta Batteries, Inc., Elmsford, NY 914–592–2500		NiMH	AAA-F	300–8,000	1.2	0.80+

Note: *Li-ion* Lithium-ion, *LiMnO*^{2} Lithium manganese dioxide, *NiCd* Nickel-cadmium, *NiMH* Nickel-metal hydride, *RAM* Rechargeable alkaline manganese

Table A.24 Electronic ceramics (between 25 and 100°C)

	96% Alumina (Al_2O_3)	Beryllia (BeO)	Boron nitrade (BN)	Aluminum nitrade (AlN)	Silicon carbide (SiC)	Silicon (Si)
Hardness, Knopp (kg/mm^2)	2,000	1,000	280	1,200	2,800	–
Flexural Strength (10^5 N/m^2)	3.0	1.7–2.4	0.8	4.9	4.4	–
Thermal conductivity (W/(m K))	21	250	60	170–200	70	150
Thermal expansion (10^{-6}/K)	7.1	8.8	0.0	4.1	3.8	3.8
Dielectric strength (k V/mm)	8.3	19.7	37.4	14.0	15.4	–
Dielectric loss (10^{-4} tan delta at 1MHz)	3–5	4–7	4	5–10	500	–
Dielectric constant, κ (at 10 MHz)	10	7.0	4.0	8.8	40	–

Table A.25 Properties of glasses

	Soda-lime	Boro-silicate	Lead glass	Alumo-silicate	Fused silica	96% Silica
Modulus of elasticity (10^6 psi)	10.2	9.0	8.5–9.0	12.5–12.7	10.5	9.8
Softening temperature (°F)	1,285	1,510	932–1,160	1,666–1,679	2,876	2,786
Coefficient of thermal expansion (10^{-6} in/in °C)	8.5–9.4	3.2–3.4	9–12.6	4.1–4.7	0.56	0.76
Thermal conductivity (BTU—in/h ft^2 °F)	7.0	7.8	5.2	9.0	9.3	10.0
Density (lb/in^3)	0.089	0.081	0.103–0.126	0.091–0.095	0.079	0.079
Electrical resistivity (Log10Ω cm)	12.4	14	17	17	17	17
Refractive index	1.525	1.473	1.540–1.560	1.530–1.547	1.459	1.458

Table A.26 Comparison of IR Transmitting Glasses Produced by AMI

Property	AMTIR-1	AMTIR-2	AMTIR-3	AMTIR-4	AMTIR-5	AMTIR-6	C1
Composition	Ge-As-Se	As-Se	Ge-Sb-Se	As-Se	As-Se	As-S	As-Se-Te
Transmission Range μm	0.7–12	1.0–14	1.0–12	1.0–12	1.0–12	0.6–8	1.2–14
Ref Index @ 10μm	2.4981	2.7613	2.6027	2.6431	2.7398	2.3807	2.8051
DN/DT°C $\times 10^{-6}$ @ 10μm	72	5	91	−23	<1	<1(5μm)	31
Knoop Hardness	170	110	150	84	87	109	110
Therm Exp $\times 10^{-6}$/°C	12	22.4	14	27	23.7	21.6	23
Thermal Condx (cal/gm sec°C) 10^{-4}	6	5.3	5.3	5.3	5.7	8.0	5.2
Specific Heat (cal/gm °C)	0.072	0.068	0.066	0.086	0.076	0.081	0.062
Density gm/cm^3	4.4	4.66	4.67	4.49	4.51	3.2	4.69
Rupture Mod (psi)	2700	2500	2500	2358	2400	2400	2500
Young's Mod ($\times 10^6$ psi)	3.2	5.6	3.1	2.2	2.56	2.3	1.8
Shear Mod ($\times 10^6$ psi)	1.3	1.03	1.2	0.85	1.01	0.94	1.03
Poisson's Ratio	0.27	0.29	0.26	0.297	0.279	0.24	0.29
Softening Point °C	405	188	295	131	170	210	154
Glass Trans Temp (Tg °C)	368	167	278	103	143	187	133
Upper Use Temp °C	300	150	250	90	130	150	120
Dispersion Values							
3 − 5 μm	202	171	159	186	175	155	148
8 − 12 μm	109	149	110	235	172		196

Index

A
α-particles, 426, 513
A/D, 5, 14, 26–28, 196–208, 210, 211
aberrations, 160, 161, 163
ablation sensor, 320–321
ABS, 613, 641, 646
absolute sensor, 8, 111, 403, 526, 549
absolute temperature, 20, 111, 127, 447, 526, 528, 530, 532, 533, 549, 553
absolute zero, 8, 118, 129, 267, 526, 557, 614
absorbent coating, 591
absorber, 130, 141, 272, 319, 482, 484
absorption, 112, 132, 154, 163, 170–172, 235, 477, 572, 595, 596
absorption coefficient, 154
absorptivity, 95, 129, 154, 158, 170, 270, 445, 482, 484, 487, 621
acceleration, 140, 327–351, 353–354
accelerometer, 140, 329–332, 335–336, 348–349, 436
accuracy, 20–21, 25, 31–34, 36–37, 180, 390, 447, 532, 537
acoustic, 10, 44, 93, 94, 114, 247, 369, 431, 434–437, 441, 587, 588
acoustic noise, 113, 249
acoustic pressure, 114, 432, 434
acoustic sensors, 431–443, 565, 587, 590
acoustic wave devices, 587
acrylic, 134, 613, 616, 641, 645, 648
active bridge, 218
active sensor, 7–8, 41, 86, 188, 249, 256, 314
actuator, 3–7
additive noise, 227, 228
AFIR, 494–497
aging, 45, 46
air bubble, 350
airflow, 392, 411–413, 421
aluminum, 434, 452, 615, 645–648
aluminum coatings, 615
aluminum nitride, 93
aluminum oxide, 453
AM, 316
Ampere's law, 71
amperometric devices, 579–581
amplifier, 178–186, 207–208, 468–470
AMTIR, 617, 620, 648, 652
analog-to-digital converter (A/D), 196–211
analyte, 572, 586
angular displacement, 290, 299, 342, 350
angular encoding, 296
antenna, 194, 250–253, 317, 321
aperture, 133–135, 157, 167, 169, 170, 316, 497, 498
appliances, 9, 247, 261
approximations, 15
array, 598
arsenic trisulfate, 617
ASIC, 179, 197, 206, 486
attenuation coefficient, 154, 318, 440
auxiliary electrode, 576
avalanche, 500, 506, 507, 509, 511, 516

B
band gap, 462, 463, 475, 515
band-gap references, 192
bandwidth, 44, 128, 224, 225, 317, 464
barometer, 375, 379
battery, 178, 243–246, 465, 649–651
bead-type thermistors, 539–540

becquerel, 504, 640
Beer's law, 132
Beer–Lambert law, 595
Bell, A.G., 115
bellows, 355, 379–381, 386–388
Bernoulli, D., 375, 392, 402
beryllium, 615
bias current, 176, 179–181, 225, 489
bias resistor, 259, 268, 269, 273, 421, 488–490
bimetal, 119–120, 349
binary codes, 197, 198, 213
biochemical sensors, 597
biological sensors, 440
bismuth, 484, 642, 644, 645
blackbody, 130, 133, 135
bolometer, 491–494, 497
Boltzmann constant, 111, 224, 467, 559
bonding, 384, 633–634
boron, 55, 609, 631, 632
Boyle, R., 375
brass, 616, 645–648
breakwire, 320, 321
breeze sensor, 420–421
bridge circuit, 106, 215–220, 242, 287, 302, 406
brightness, 154
British Unit of Heat, 117
broadband detectors, 482
bubble chamber, 517

C

cable, 177, 194, 195, 233–234, 236, 346–348, 520, 521
cadmium telluride, 515, 516
calibration, 20–25, 33, 36, 37, 242–243, 410, 532
calibration error, 34–35
calibration temperature, 119, 526
Callendar–van Dusen approximations, 528, 529
candela, 11, 155, 638
cantilever, 336, 414, 590–592
cantilever beam, 361, 632
capacitance, 60–67, 89, 209, 231, 254, 284–285, 438, 451, 468, 579, 583
capacitance-to-voltage converter, 208–210, 257, 333
capacitive accelerometer, 332–334
capacitive bridge, 286, 287
capacitive coupling, 231, 256–258, 367
capacitive sensor, 183, 208, 284–288, 322, 366, 387–388, 448–451
capacitor, 61–65, 87, 178, 183, 192, 195, 200–202, 208–210, 234–235, 255–256, 284, 332–333, 365–366, 449–450, 470, 582
catalytic devices, 594–595, 602
cavity, 25, 306, 307, 338, 391, 493, 521
cavity effect, 133–135
CCD, 479
CdS, 473, 474
Celsius, A., 117, 118, 447, 527, 560
ceramic, 83, 89–94, 96, 178, 214, 315, 347, 436, 539–541, 617, 642
characteristic temperature, 83, 535, 536
charge, 209
charge amplifier, 183–186
charge detector, 256
charge-balance, 200–202
charge-to-voltage converter, 183, 209, 583
chemFET, 584–585
chemical poisoning, 602
chemical reaction, 576, 593, 594, 603
chemical sensor, 3, 569–603
chemical species, 581, 584
chemicacpacitive sensors, 582
chemiresistor, 581
chip thermistor, 540, 541
circuit protection, 548
cladding, 166, 167, 241, 596
Clark electrode, 579, 580
clock, 209
cloud chamber, 516
CMOS, 478, 480–481, 489
CMRR, 179, 182, 229
CO, 477, 577
CO_2, 477, 577
coating, 86, 159, 160, 171, 490, 584, 589–591, 615–617
coaxial cables, 194–195, 235, 286
cobalt, 73, 563
coefficient of reflection, 129, 150, 158, 620
coil, 71, 74–77, 288, 290–292, 302, 328, 329, 346, 388, 390, 439–440, 455, 494
cold junction, 109, 338, 483, 484, 552–553
collector, 471–472, 559
comparator, 192, 193, 200–205, 259, 262
complex devices, 633
complex sensor, 3, 262–263, 377
concentrator, 169–170
condenser microphones, 432–433
conductance, 414, 473–474, 578, 585
conduction, 78, 121–125, 141, 463, 473
conduction band, 463, 473
conductive plastics, 614
conductivity, 80, 291, 318, 462, 463, 575, 578

conductivity sensor, 452–456
conductometric devices, 452, 573, 578–579
constantan, 550, 643
contact resistance, 123, 124
contact sensor, 525, 526
convection, 121, 122, 125, 337
copolymer, 95–96, 614
copper, 54, 55, 78, 241, 347, 434, 435, 616, 621, 626
Coriolis, G.-G., 342, 422–423
Coriolis acceleration, 342, 344
Coriolis force, 342, 422
Coriolis tube, 376, 422
Coulomb's law, 56
coupling (thermal), 519
cross-talk, 500
crystal, 87–90, 97, 335, 371, 372, 458, 462, 473, 487, 565, 587, 620, 630, 646
Curie point, 98, 546
Curie temperature, 49, 73, 90, 91, 100, 101, 292, 347, 545, 549
current, 575–581, 585, 586, 592, 594
current generator, 188–191, 454
current mirror, 190, 559
current sink, 189
current source, 189, 191, 200–202, 221, 468
current-to-voltage (I/V) converter, 183, 186–188, 584–585
CVD, 93, 451, 624

D
damping, 44, 45, 139–141, 330
Darlington connection, 472
data acquisition, 1–12, 26
dead band, 38, 39, 294
decibel, 30, 31, 115
definition (sensor), 2
deflection, 119–120, 216, 379–381, 388–390, 433–435
dew point, 447, 456–458
Dewar, J., 323, 476
Dewar cooling, 475
diaphragm, 379, 381, 383–391, 433–435, 437–438, 440, 631
dielectric, 87, 365, 582
dielectric absorption, 178, 234–235
dielectric constant, 63–67, 88, 323, 324, 448, 450, 641
differential equation, 41–43, 141, 142, 495, 523, 543, 544
differential sensor, 229, 269, 274, 386, 404
diffusion, 579–581, 631
digital format, 5, 39, 220, 351, 405
digital-to-analog converter (D/A), 203–206
diode, 186, 187, 250, 251, 465, 467–469, 558
dipole, 58–59, 63, 90–91, 98, 99
direct conversion, 54, 109, 208, 335, 466, 467
direct sensor, 4, 53, 96
disbalanced bridge, 216–218
displacement, 15, 32, 114, 140, 279–325, 327–328, 332, 377, 379, 387, 390
displacement sensor, 63, 168, 277, 279, 280, 284–286, 309, 312, 313
dissipation constant, 546, 549
dissipation factor, 543
distance sensor, 297
distortion mask, 263
divider, 185, 212–214
door openers, 252, 253
Doppler, C.J., 250–253, 417, 418
Doppler effect, 250, 251, 314, 416
drag element, 423
drag force sensor, 423–424
driven shield, 177, 256, 257, 286, 287
drivers, optical, 196
dual-ramp, 197
dual-slope, 203
dust detector, 424
dynamic error, 41
dynamic microphone, 439
dynamic range, 30, 334, 347, 361, 588
dynodes, 505–507

E
e-nose, 599, 600
eddy currents, 290–292
Einstein, A., 68, 186, 251, 348, 461
elasticity, 113, 367, 615
electret microphone, 437–439
electric charge, 54–61, 68, 70, 77, 87, 99, 103, 126, 258, 271, 432
electric current, 8, 47, 53, 68, 69, 73, 77–79, 104, 112, 251, 506, 508
electric dipole, 58–59, 98
electric field, 55–60, 63, 77, 78, 91, 137, 257–260, 508
electric potential, 12, 60, 104
electrical conduction, 78, 103
electrochemical cell, 577, 578
electrochemical sensor, 575–578
electrode, 85–88, 98, 257–259, 268, 269, 286, 287, 311, 323, 350, 367–368, 419, 433, 441, 449, 451, 452, 454, 474, 508, 576, 577, 579–581, 589
electrolyte, 244, 350, 576–580, 602
electromagnetic flowmeter, 419, 420

electromagnetic radiation, 116, 126, 130, 138, 260, 267, 461
electromagnetic sensor, 418–420
electrometer, 577, 585
electromotive force, 73–74, 107–108, 240–241, 417, 419
electron, 54, 68–69, 77, 78, 103, 104, 108–109, 426, 462, 463, 466, 473, 500, 506–508, 512, 577
electron multiplication, 500, 507
electron–hole pairs, 465, 512, 513
electronic nose, 569, 599–602
electroplating, 625
electrostatic, 231, 285
electrostatic gyro, 341
electrostatic shield, 213, 232
emissivity, 129–135, 171–172, 267, 648
emitter, 130, 196, 272, 426–428, 471, 472, 559
encoding disk, 309–310
energy bonds, 611
enzyme, 580–581, 585, 593, 597, 598
enzyme sensors, 597–598
epoxy, 334, 540, 541, 614
etching, 384, 609, 628–633
Euler, L., 353
excimer laser, 627
excitation, 7, 41, 188–196, 220–222, 249, 283, 289, 420, 441, 463, 561

F

Fabry–Perot sensors, 306–308, 390
Fahrenheit, G., 117, 118
far-infrared 130, 131, 135, 136, 153, 159, 164, 170–172, 267–274, 494–497, 619–620
farad, 61
Faraday cage, 58
Faraday, M., 63, 69, 73, 418, 577, 580
Faraday's Law, 69–71, 73, 74, 328
feedback, 180, 190, 193, 195, 406, 412, 413, 433, 468
Ferdinand II, 117
Fermi, E., 111
ferroelectric, 86, 91, 98, 100
ferromagnetic, 69, 288, 292, 293, 295
FET, 225
fiber, 166–168, 304–305, 346, 434, 562–563, 596, 614, 645
fiber-optic, 165–169, 346, 434–435
fiber-optic sensor, 167–168, 304–305, 596–597
field lines, 56, 68, 70
filament, 81, 396, 547
film, 452, 493
film transducers, 96
filter (optical), 137, 148, 178, 303, 304
first-order response, 42, 44
flame, 498, 500
flow, 417–418
flow measurement, 401, 402
flow rate, 34, 141, 392, 401, 404, 406, 419, 420
flow resistance, 79, 402
flowmeter, 392, 393, 404, 417–420, 422, 423
fluid, 1, 21, 22, 56, 125, 355, 375–376, 418, 422–424
fluoroplastics, 613
fluoroptic method, 562
flux, 3, 56, 72, 74, 129, 133–134, 150–156, 272, 288, 297, 401, 498
focus, 162, 163
focusing lens, 261, 265, 302, 498
foil, 556
follower (voltage), 181
force, 55, 353–365, 423–424
forced convection, 125
Fourier, J., 112
FP interferometer, 390
FPA, 493
Fraden model, (thermistor), 535–537, 539
Franklin, B., 54
frequency range, 44, 64, 249, 431
frequency response, 42, 175, 331, 472
Fresnel, A., 163
Fresnel lens, 163–165, 265, 270, 612
frost point, 456
FSR, 362–364
full scale, 30–31, 33–34, 197–199, 206, 223, 308

G

γ-radiation, 503, 504, 505, 512, 514, 515, 616
Gain–bandwidth product, 180, 472
Galileo, 353
gas analyzer, 477
gas chromatographs, 583
gas sensor, 584, 589, 594, 596
gauge, 84, 85, 290, 334, 335, 355–357, 393–396
gauge sensor, 386
Gauss' law, 56, 58
Gaussian System, 11
Geiger, J. W., 510
Geiger–Müller counter, 510–512
geometrical optics, 147
geometry factor, 62, 63, 75, 80, 85
germanium, 152, 493, 494, 515, 541
Gilbert, W., 67

glass, 163, 166, 390, 540, 541, 617–620, 633, 634
glucose, 580, 581, 593
Golay cells, 482–483
gold, 159, 171, 616, 625, 626, 629
gold black, 648
GPR, 318
GPS, 327, 339
grating, 308–310
gravimetric detector, 588
gravitational sensor, 348–351
gravity, 331, 337, 348, 349
ground penetrating radar, 318
Gunn oscillator, 250
gyroscope, 325, 339–346

H
H_2O, 132, 278, 448, 594
Hall, E.H., 103
Hall effect, 103–106, 383, 608
Hall effect sensor, 103–106, 293–297
harmonic, 49, 114, 248, 251, 371, 418, 439, 590
heat, 519
heat absorption, 98, 112, 543
heat capacity, 120–122
heat loss, 394, 407, 494–496, 543, 547, 593
heat pump, 456–458
heat sink, 102, 242, 336, 337
heat transfer, 121–135, 336, 394, 491, 495
heated probe, 598
Henry, J., 73, 75, 418
Hooke, R., 117
hot junction, 109, 111, 338, 483–485
hot-wire anemometer, 405–408
Howland, B., 190, 191
humidity, 19, 48, 49, 66, 86, 93, 445–448, 602
humidity sensor, 445–458
hydrocarbon fuel, 581
hydrocarbon sensor, 602
hydrogel, 585, 598
hydrophone, 309, 431, 437
hygristor, 85, 86, 216
hygrometer, 25, 445, 456–458
hygrometric sensor, 399, 452
hysteresis, 35–36, 290, 293, 295, 364, 381, 456, 457

I
illumination, 147, 149, 155, 187, 324, 475
image, 162, 248, 261–266, 269, 271, 303, 325, 478–481, 628
immobilization (biochemical), 597, 598
inclination detectors, 6, 349
index of refraction, 150, 159, 166, 505
inductance, 75–77, 234, 238, 239, 292, 389, 390, 638, 639
inertia, 43, 44, 140, 341, 343, 353
inertial mass, 332, 335, 338
Infrared (IR), 128, 152, 260, 491, 561
infrared flux, 3, 133, 272, 481
infrared (IR) sensor, 6, 39, 131, 132, 248, 476, 540, 621, 643
infrasonic, 431, 439
inherent noise, 207, 223–227
input impedance, 40, 174–176, 221, 436, 439, 454, 490
input resistance, 99, 174, 181, 182
input stage, 175–178, 239, 328
instrumentation amplifier, 182, 200
insulation, 46, 142, 257, 347, 494, 555, 613
integrator, 200–202, 208, 234
intensity sensor (optical), 168
interface circuit, 19, 97, 173–246, 415, 428
interferometer, 345, 434, 483
intrusion, 248, 255, 256, 260, 276, 277
ion, 383, 508–512, 578, 585, 591–593, 632, 650, 651
ionization, 395–396, 424–426, 507–516, 586
ionizing chamber, 508
ionizing radiation, 46, 188, 503, 504, 510, 514, 516
iridium, 616, 637, 642, 645
ISA, 549
ITS-90, 527, 529

J
JFET, 239, 269, 421, 439, 488, 489
Johnson noise, 224, 470
Joule, J., 111, 112, 483
Joule heat, 78, 112, 412, 542, 547
junction, 23, 109, 112, 143, 186, 241, 338, 339, 465, 483, 484, 513, 549, 552, 553, 556–558, 643, 644
junction capacitance, 465, 470

K
Kawai, H., 93, 98
Kelvin, 8, 117, 118, 271, 273, 448
keyboard, 354, 357, 361
kinetic energy, 108, 116, 120, 126, 127, 171, 376, 462, 506, 623
Kirchhoff, G.R., 80, 129, 170
Kirchhoff's laws, 80, 142, 143
KOH etch, 630–632
Korotkoff sounds, 436

krypton, 512, 586
KTY, 530, 531

L
Laplace transforms, 329, 330
laser, 137, 148, 248, 249, 434, 591, 627, 648
laser gyro, 344–346
LC, 565
LCD, 137, 211, 368
lead, 89, 100, 234, 245, 335, 472, 616, 618, 637, 642, 645–647
leakage current, 176–177, 465, 513–516
least squares, 25, 51
LeChatelier, H., 110
LED, 196, 218, 274, 276, 310, 350, 369, 563
lens, 161–165, 260–267, 270, 272, 273, 498
Leslie, J., 445
level detectors, 294, 295, 304, 305, 549
life test, 48, 49, 244
LIGA, 627, 634–635
light, 135–138, 147–154, 163, 166–168, 186–188, 196, 261, 264–267, 274, 275, 303, 305–307, 309, 310, 345, 427
light polarization, 136, 303–304
light scattering, 137, 426–427
light-to-voltage converter, 186, 473, 467–470, 475
linearity, 21, 36, 37, 331, 387
liquid, 90, 121, 294, 295, 304–305, 322–324, 446, 517
lithium, 245, 514–515, 637, 649
load cells, 354, 386
logarithmic scale, 31
logic circuits, 192, 253, 256
long-term stability, 45, 528, 541
loudspeaker, 3, 93, 113, 226, 439
lumen, 154–156
luminescence, 500, 595
LVDT, 288–290, 328, 355

M
magnesium, 448, 615, 620, 637
magnet, 67, 68, 72, 73, 294–299, 301, 328, 646
magnetic field, 67–72, 74, 103–105, 235, 236, 290, 292, 294, 297, 298, 301, 440, 638
magnetic flux, 72, 295, 296, 418, 419, 638
magnetic noise, 235
magnetic pole, 67, 294
magnetic reluctance sensor, 302
magnetic sensor, 72, 216, 288–302
magnetic shielding, 235–237
magnetism, 67–73, 639
MagnetoPot, 283
magnetoresistive sensor, 70, 297–300
magnetostrictive detector, 300–302
manganese, 73, 83, 244, 245, 637, 642
mass spectrometers, 570
material characteristic (thermistor), 534–536
matrix, 73, 86, 486, 556, 581, 593
Maxwell, J.C., 419
MCT, 476, 477
measurand, 2, 50
membrane, 276–278, 379–381, 390, 396–397, 404, 438, 482, 484, 493, 578, 579, 585, 589
MEMFET, 585
MEMS, 364–365, 583, 601, 608, 626
MEMSIC, 337
MEOMS, 626
mercury switch, 349, 350
metal carbides, 617
metal films, 171, 225, 408, 493, 622, 629, 633
metal oxide, 171, 532, 540, 572, 574–575
metallization, 159, 437, 540, 614
Michelson, A.A., 434
microbalance method, 586, 587
microbalance sensors, 440, 587, 588
microcalorimetry, 593
microcantilever, 590
microcontroller, 182, 192, 203, 207, 211, 281, 369
micromachining, 306–307, 609, 626–635
microphone, 416, 431–440, 621
microsensor, 334, 404, 415, 601
microwave, 248–254, 315, 491, 545
mid-infrared, 135, 461, 481, 613
military standard, 49, 554
MIR, 316–318
mirror, 158–161, 190, 263, 306, 390, 434, 456–458, 482, 618
modulation, 15, 40, 192, 316, 317, 388, 391, 595
moisture, 85–86, 94, 445–458, 555, 643
molybdenum, 160, 616–617, 637, 644, 645
monolithic sensors, 341, 558, 560
MOS, 209, 210, 259, 310, 574
mouse, 324
motion detector, 247–278, 487
MTBF, 47, 48
Müller, W., 510
multiplexing, 203, 212, 213, 360
multiplicative noise, 227–230, 237
multivibrator, 192, 200, 475
mutual inductance, 76, 77, 288

N
n-wells, 105, 106
natural frequency, 6, 43–44, 330, 331, 380, 587
near-infrared, 135, 136, 274, 445, 498, 516
Nernst equation, 577
neural network, 600
neuron, 599, 600, 602
Newton, I., 28–30, 117, 150, 353
Newton's law, 125, 140, 143, 327, 341, 354, 522
Newton-Raphson method, 28
nichrome, 95, 171, 356, 487, 493, 621, 624
nickel, 73, 245, 616, 626, 637, 642, 644, 645, 647, 648
noise, 52, 178, 200, 207, 209, 223–242, 248, 271, 316, 433, 464, 470, 475, 485, 507
nonlinearity, 16, 36–37, 531, 558
NTC (thermistor), 81, 83, 395, 532–533, 539–545, 574
nuclear radiation, 94, 503, 504, 555, 618
nylon, 283, 613, 641, 645–647
Nyquist, H., 206, 207, 420

O
occupancy sensors, 247
odor sensor, 599
Oersted, H.C., 68
offset, 181, 214, 224, 449, 489
offset voltage, 176, 179–181, 489
Ohm's law, 15, 62, 76, 79, 189, 269, 402, 521, 546
olfaction, 569
olfactory cells, 599, 600
one-shot, 200–202
open-loop, 179, 345
open-loop gain, 180, 469, 470
operational amplifier (OPAM), 176, 178–181, 184, 187, 189, 191, 193, 208, 210, 489
optical cavity, 306, 345, 390
optical contrast, 247, 261, 262, 266, 302
optical detection, 466, 467
optical modulation, 595
optical paths, 167, 306, 345, 596
optical power, 466, 467, 477
optical sensor, 302–314, 369–370, 497, 560–563, 595
optocoupler, 309, 457
organic, 90, 171, 511, 541, 586, 621, 623, 627
oscillating hygrometer, 458
oscillating response, 44, 470
oscillator, 192–194, 208, 250, 342, 371, 433, 587
output capacitance, 177, 181
output current, 189–191, 194, 312, 467, 509
output impedance, 40, 175, 176, 234, 454
output resistance, 105, 176, 179, 181, 189, 221, 488
output signal format, 3, 4
oversampling, 206
oxygen, 87, 569, 579–581, 637, 645

P
p-n junction, 22, 23, 186
p-substrate, 105
palladium, 576, 585, 594, 616, 637, 642, 645
parallel-plate capacitor, 61, 62, 366, 432, 582, 583
Pascal, B., 79, 94, 375, 638, 640
passive sensor, 5–7, 178, 245, 549
pellister, 594–595
Peltier, J.C.A., 111
Peltier effect, 111–113, 457, 475, 495
phase lag, 43, 175, 276, 469
phase shift, 43, 195, 324, 416, 422, 469, 562
phenolic, 613, 614, 645
phosphor, 561, 562, 616, 645, 646
photocathode, 505–507
photocurrent, 186, 187, 391, 457, 465, 470, 471
photodetector, 265, 274, 302, 457
photodiodes, 186–187, 350, 465–471
photoeffect, 4, 186, 310, 461–463
photoelectron, 479, 505–507
photoionization detector, 585
photomultiplier, 461, 505, 506
photon, 13, 53, 136, 303, 461–465, 506, 511
photoresist, 628, 629, 634
photoresistor, 265, 266, 473–475
photosensor, 186, 187, 426, 472
phototransistors, 186, 471–472
photovoltaic mode, 467, 468
piecewise approximation, 18–19, 26–28
piezoelectric, 86–96, 268, 315, 335–336, 346–348, 354, 370–372, 416, 441, 565, 581, 588, 608
piezoelectric crystal, 89, 192, 416–418, 435, 587
piezoelectric film, 93–96, 359–362, 436, 589, 621
piezoelectric hygrometer, 458
piezoelectric plastics, 614
piezoresistive accelerometer, 334–335
piezoresistive bridge, 335, 385
piezoresistive effect, 84, 335, 356, 381
piezoresistive gauge, 215, 335, 381
piezoresistive sensors, 362–364, 381–387, 390
pinhole lens, 265

pink noise, 225
pipe, 77–79, 113, 275, 400, 401, 409
PID, 585–586
PIR, 267, 272–274, 483, 487, 494, 496
Pirani gauge, 393–395, 545
Planck, M.K.E.L., 127
Planck's constant, 136, 461
Planck's law, 127
plano-convex lens, 162
plastic, 73, 166, 270, 309, 611–615, 619, 642, 645
platinum, 82, 83, 171, 407, 408, 493, 528, 540, 594, 602, 616, 637, 642, 644–647
pointing devices, 324
Poisson ratio, 114
polarization, 10, 63, 91, 100, 136–137, 303, 304, 579, 595
polarization filter, 137
poling, 59, 90, 91, 98, 421
polycarbonate, 275, 613, 646
polyester, 283, 613, 614, 641, 645, 647
polyethylene, 130, 165, 612, 613, 641, 645, 646
polymer, 93–96, 269, 347, 363, 449, 581, 583, 611, 612, 614, 628, 643
polymer matrix, 581
polymerization, 612–614
polypropylene, 358, 613, 619, 641
polysilicon (PS), 493, 609, 610, 629
polystyrene, 452, 585, 611, 613, 649
polyurethane, 363, 585, 613, 647
popcorn noise, 224, 226
position, 114, 279–325, 327
position-sensitive detector, 482
potential, 54–61, 104, 109, 238, 365, 506, 551, 574, 577, 586, 639
potentiometer, 220, 281–284
potentiometric devices, 38, 280–284, 573, 576–578
preaging, 46, 541
predictive, 493, 519, 520, 524, 525
presence sensor, 254, 274
pressure, 354, 375–397
pressure gradient, 276–278, 392, 402–404, 440
pressure sensor, 8, 247, 276–278, 307, 355, 375–397, 403
primary cells, 243–245, 649
prototype, 46, 238, 239, 321
proximity, 256, 279, 366, 434
proximity sensor, 279, 286–288, 292, 303–304, 369, 388
PSD, 307, 310–314, 482
PTC thermistor, 545–549
PVDF, 93–96, 270, 347, 359–362, 643
PWM, 15, 197, 412, 413
pyroelectric, 96–102, 268, 270, 487–491
pyroelectric coefficient, 93, 100, 269, 273
pyroelectric current, 214, 273, 489, 490
pyroelectric sensor, 97–100, 102, 268, 487–491
pyrometry, 481
PZT, 93, 94, 335, 441, 643

Q
Q-spoilers, 195
quantification, 569, 571
quantum detector, 136, 461, 475
quartz, 86–87, 371, 372, 408, 458, 565, 590
quenching, 511

R
radar, 249, 254, 316–320
radiation, 116, 121, 136, 261, 508, 512–513, 519
radiation bandwidth, 128
radiation detection, 508
radiation spectrum, 126, 136
radio waves, 318, 319
radio-frequency (RF), 93, 194, 231, 433
radioactivity, 503, 504
ratiometric technique, 211, 213, 215, 229, 230, 312
Rayleigh waves, 588
reactive ion etching, 632
redox reactions, 577
reference diode, 391
reference electrode, 576
reference sensor, 25, 229, 553
reference temperature, 119, 549, 553
reference voltage, 219, 243, 412
reflection, 134, 137, 148–152, 158, 159, 169, 274
reflective surface, 160, 168
reflectivity, 129, 158, 457
refraction, 148, 150, 165
refractive index, 131, 150, 151, 153, 157, 165, 166, 305, 618, 620
relative humidity (RH), 45, 48, 49, 66, 67, 445–449
relative sensors, 8, 526
reliability, 47–49, 248, 360
repeatability, 38, 39
resistance, 77–86, 280, 310–311, 378, 610
resistance multiplication, 185–186
resistive bridge, 106, 210, 216, 219, 242
resistive load, 189, 191, 468
resistive sensor, 218, 219

resistivity, 79–81, 452, 484, 485, 513, 516, 530, 609
resistor, 79, 185, 200–201, 212, 221, 269, 381, 387, 414, 421, 432, 545
resolution, 38–39, 198, 199, 205–207, 223, 281, 282, 342, 505
resonant, 44, 321, 330, 331, 438, 565, 587
resonant sensors, 632
resonator, 343–346, 371, 372
retina, 170
rhodium, 575, 616
roentgen, 504
root-sum-of-squares, 51
rotor, 340–341
RTD, 82, 83, 410, 528–530, 545, 546
RVDT, 288–290

S
Sagnac effect, 344, 345
sampling rate, 207
saturation, 37, 38, 446, 447, 509
SAW, 369, 441, 442, 587–590
scale, 8, 31, 117–118, 527
Schmitt trigger, 300
Schottky noise, 225
scintillation counter, 505
second-order response, 43, 44
secondary cells, 245–246, 649, 650
security alarms, 252
security system, 255
Seebeck coefficient, 107, 109, 111
Seebeck effect, 106–112, 608
Seebeck potential, 108, 109, 551, 552
Seebeck, T.J., 106, 110
selectivity, 571, 597
self-heating, 47, 395, 533, 410, 542–547
self-induction, 74, 292
semiconductor detectors (ionizaion), 512–516
semiconductor diode, 512, 513
sensitivity, 17–18, 84–86, 216–218, 315, 330–331, 339, 356–358, 382–383, 433, 438–439, 474, 534, 553, 571, 572, 600, 610, 644
sensor classification, 7
sensor definition, 2
SFB, 384
shield, 177, 232–236, 257, 286, 521–522
shielding, 230–234, 436
signal conditioning, 173, 210
signal-to-noise ratio, 417, 435
smoke detector, 324–328
signatures, 369, 581
silicate, 607, 617
silicon, 55, 105, 310–311, 332–333, 343, 381, 386, 390, 451, 529–531, 557, 584, 589, 607–611
silicon bonding, 628, 634
silicon diaphragms, 381, 382, 385, 387, 630–631
silicon diode, 490, 513
silicon dioxide (SiO_2), 384, 415, 451, 627
silicon micromachining, 628–635
silicon nitride, 582
silicon plate, 404, 589
silicon sensor, 105, 530, 531
silicon wafer, 627–631, 633, 634
silicone, 359, 581
silicone oil, 330
silkscreen, 85, 358, 540
silver, 90, 616
skin, 134, 135, 238, 358, 360
Snell, W., 150
Snell's law, 150, 151, 162, 165, 167
SnO_2, 575, 601
soil, moisture sensor, 453–455
solder, 48, 241, 551
solenoid, 71–76
solid-state detectors, 512, 513, 516
sorbent coating, 590
sound waves, 113–116, 369, 431, 646
span errors, 19, 32, 34
species, 571, 574, 581, 584, 591, 597, 602
specific heat, 113, 120, 121, 647
specific resistivity, 79–81, 85, 609
spectrometry, 591, 592
spectroscopy, 509, 514, 515
spectrum, 113, 126, 136, 153, 371, 498, 503
speed, 113, 114, 126, 197, 251, 315, 327, 423, 564, 646
speed response, 42, 187, 296
spherical mirrors, 160, 161
spin-casting, 621–623
spinning-rotor gauge, 396
spline interpolation, 19
sputtering, 93, 623–624
square-wave oscillator, 192, 193, 200, 257
statistical methods, 50, 51
Stefan–Boltzmann constant, 129, 272, 415
Stefan–Boltzmann law, 128–129, 267, 271, 414, 485, 495
Steinhart and Hart model, 537–539
stimulus, 2, 7, 10, 13–15, 26–31, 33, 38, 41–43, 50, 138, 167, 212, 228, 229, 440, 442
storage, 42, 45, 49, 96
straight line, 18, 25, 36, 37, 82

strain, 84–85, 88, 334, 353–372, 379, 423–424, 608, 610
strain gauge, 84, 168, 169, 334, 355–357, 383
stress, 84, 91, 98, 268, 371, 380–381, 383, 387, 388, 591
stress detectors, 247
string, 113, 372
substrate, 83, 86, 283, 361, 408, 411, 441, 451, 493, 582, 588, 617, 624, 626, 627
successive-approximation (A/D), 197
switched-capacitor amplifier, 209
synchronous detector, 289
systematic error, 32, 598

T
tactile sensor, 353–372, 632
target species, 571
TCR, 80, 415, 490, 492, 545, 546, 642
Teflon, 177, 438, 555, 580, 613, 641, 647
temperature, 46, 64, 78, 96, 97, 108, 116–121, 153, 192, 224, 267, 336, 403, 481, 544
temperature coefficient, 81, 383, 395, 471, 530, 610, 642, 644
temperature compensation, 222, 372, 383, 387, 405, 530, 558
temperature correction, 66, 387
temperature differential, 405, 484, 492, 552
temperature gradient, 107, 125, 142, 338, 407, 410, 517, 521, 522, 543
temperature profile, 90, 122–124, 524
temperature sensitivity, 80–83, 302, 381, 546, 565, 608, 610
temperature sensor, 338, 339, 387, 409–411, 481, 519–566
TGS, 49, 98, 490, 491, 643
thermal accelerometer, 336–339
thermal capacitance, 102, 521
thermal conductivity, 122–124, 393, 394, 408, 455–456, 484, 525, 638, 639, 643, 646–648
thermal coupling, 482, 491, 493, 520, 549, 553
thermal expansion, 118–120, 268, 344, 618, 646, 651
thermal feedback, 549
thermal flux, 130, 135, 271, 273, 485, 495, 496
thermal grease, 125, 521, 548
thermal mass, 483, 496
thermal radiation, 126–135, 267, 407, 481, 526, 619, 621
thermal resistance, 102, 123, 141, 521, 533, 543, 639
thermal shock, 49, 618
thermal time constant, 100, 102, 523, 524, 546, 556, 601
thermistor, 8, 21, 22, 27, 28, 47, 81–83, 212, 395, 412, 455, 532–549, 554, 593
thermoanemometer, 404, 405, 409–414
thermochromic solution, 563
thermocouple, 8, 109, 110, 338, 483, 549–556, 643, 644
thermocouple amplifier, 553
thermocouple assemblies, 554–556
thermodynamics, 11, 116, 141, 593
thermoelectric, 106–113, 241, 484, 485, 549–556, 644
thermoelectric coolers, 112, 475
thermoelectricity, 108
thermometer, 116, 117, 486
thermopile, 3, 111, 338, 483–487, 552, 644
thermoplastic, 612–614
thermostat, 135, 349, 548, 549
thermowell, 554
thick film, 83, 96, 411, 621, 641
thin film, 93, 322–323, 493, 581, 608, 621, 628–629
thin plate, 379–381, 388
Thompson, W., 107
Thomson heating, 112
threshold, 105, 192, 200, 208, 259, 262, 293, 350, 354, 357, 463, 509
tilt sensor, 350
time constant, 42, 43, 102, 142, 256, 490, 523, 524, 561
TiO_2, 575, 601
titanium, 83, 92, 93, 415, 601, 637, 641, 642, 647
toroid, 71, 72, 76
torque, 59, 140, 341, 638, 639
Torricelli, E., 375, 377
total internal reflection, 160, 166–168, 305
touch screen, 367–369
touch sensor, 357–370
transceiver, 250, 253
transfer function, 13–26, 28, 29, 32–34, 36, 37, 187, 531, 560
transistor, 189, 210, 471, 557–559, 584
transition temperature, 546, 548
transmission, 3, 220–222, 254, 315, 323, 588, 652
transmittance, 152, 153, 157, 158, 615
transmitted noise, 51, 178, 227–231, 248
triboelectric detectors, 248, 258–260
triboelectric effect, 46, 54, 258
tube, 1, 169, 350, 378, 392, 399, 409, 419, 422, 506, 564, 621

tube of flow, 399, 400, 402, 416, 419
tungsten, 81, 407, 617, 622, 637, 642, 644, 647
two-point calibration, 22, 25, 34, 534
two-wire transmitter, 220

U
U.S. Customary System, 12, 354, 377, 401
ultrasonic, 249, 254, 314–316, 416–418, 564, 565
ultrasonic waves, 301, 314–316, 369, 416, 417
ultraviolet (UV), 135, 561, 585–586, 628
uncertainty, 25, 50–52, 282

V
V/F, 198–203
vacuum, 62, 136, 150, 341, 364–365, 376, 550
vacuum chamber, 622, 623
vacuum deposition, 622–623
vacuum sensor, 393–397
vacuum tube, 395, 396, 506
valence band, 463, 473
VCR, 490, 491
vehicle, 6–7, 297, 349
velocity, 70, 114, 139, 150, 252, 327–351, 399–402, 410, 416, 418, 419
velocity of light, 136, 150
velocity sensor, 328, 329
vibrating gyro, 342
vibration, 49, 120, 237, 248, 253, 329, 335, 343, 344, 369
vibration detectors, 248
virtual ground, 184, 190, 489
Volta, A., 243
voltage follower, 181, 488–490
voltage offset, 181, 224
voltage source, 105, 175, 196, 219, 542, 547, 552
voltage-to-current converter, 220
voltage-to-frequency (V/F) converter, 198–203
Voltaic pile, 68, 625
voltammetry, 577
VRP, 388–390

W
Warburg impedance, 579
warm-up time, 41
warping, 119, 120
water, 436
water tank, 65, 120
water-level sensor, 65, 66
waveguide, 165–169, 301, 302, 321
wavelength, 94, 127, 131, 132, 135–136, 138, 150, 155, 157, 165, 249, 307, 345, 391, 435, 463, 465, 473, 475, 481, 595, 596, 619
weber, 72
Wheatstone bridge, 215, 216, 356, 383, 390, 411, 575, 578
white noise, 225, 316
Wiedemann effect, 301
Wien's law, 127, 128
window, 157–158, 276, 481
window comparator, 259, 262
Winston cone, 170
wiper, 280–284
wire, 68, 71, 84, 220–222, 236–239, 280–282, 320, 356, 405–410, 512, 556
work function, 462
working electrode, 576, 577

X
X-ray, 503, 570, 615, 634
xenon, 512, 561, 586

Y
Young's modulus, 88, 114, 608

Z
zinc, 244, 617
zinc oxide, 92–93, 575, 608